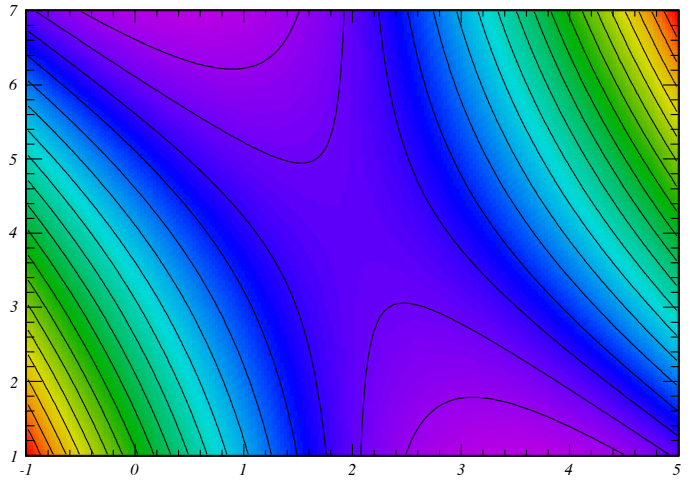
Statistical Methods

– things to remember when doing precision measurements –

Michael Schmelling – MPI for Nuclear Physics

Selected Topics:

- *Introduction*
- *Uncertainties and Error Propagation*
- *Least Squares for Professionals*



Literature

➜ *a few selected books in alphabetical order. . .*

- R.J. Barlow,*Statistics*, Wiley
- S. Brand, *Data Analysis*, Springer
- G.D. Cowan,*Statistical Data Analysis*, Oxford University Press
- H.L. Harney, *Bayesian Inference*, Springer
- F. James, *Statistical Methods in Experimental Physics*, World Scientific
- D.E. Knuth, *The Art of Computer Programming*, Addison Wesley
- W.T. Press et al., *Numerical Recipes*, Cambridge University Press
- D.S. Sivia, *Data Analysis - A Bayesian Tutorial*, Oxford University Press

➜ *What are statistical methods?*

- recipes for data reduction: large data set ➜ single number e.g.. . .
  - ➜ md5sum: fingerprint characterizing the data set
  - ➜ arithmetic average: estimate of a common underlying value
  - ➜ standard deviation: measure of uncertainty
- statistical method are constructed
  - ➜ neither “right” nor “wrong” – characterized by properties
  - ➜ properties of a method need to be understood
    to judge the applicability and to interpret the results
  - ➜ alternative estimates of a common underlying value:
    - weighted average or median
  - ➜ alternative estimates of uncertainty:
    - smallest 68% quantile or full width at half maximum (FWHM)

Mean value, standard deviation & variance

A measure for the scatter s of x with PDF $f(x)$ around a point a is:

$$s^2 = \int dx\,(x-a)^2 f(x)$$

For s to characterize $f(x)$, a should be chosen to minimize s:

$$\frac{\partial s^2}{\partial a} = -2\int dx\,(x-a)f(x) \stackrel{!}{=} 0 \quad \text{i.e.} \quad a = \int dx\; x\, f(x) = \langle x \rangle$$

The mean value $\langle x \rangle$ is the location parameter that minimizes the scatter s.

➜ *note:*

- for symmetric PDFs $\langle x \rangle$ is the symmetry point
- the scatter around $\langle x \rangle$ is called "standard deviation" σ
- σ^2 is the "variance" of a PDF

Expectation values

Given a PDF $f(x)$ and a function $a(x)$, the expectation value $\langle a \rangle$ is:

$$\langle a \rangle = \int_{-\infty}^{\infty} dx \; a(x) f(x)$$

- mapping of functions $f(x)$ to a real numbers

➜ *examples:*

$$\langle x \rangle \quad : \quad \text{mean value}$$

$$\left\langle (x - \langle x \rangle)^2 \right\rangle : \quad \text{variance}$$

- important property: linearity i.e. $\langle \alpha A + \beta B \rangle = \alpha \langle A \rangle + \beta \langle B \rangle$

➜ *application e.g.:*

$$\left\langle (x - \langle x \rangle)^2 \right\rangle = \left\langle x^2 - 2x \langle x \rangle + \langle x \rangle^2 \right\rangle = \left\langle x^2 \right\rangle - \langle x \rangle^2$$

➜ *what are "uncertainties"?*

- measures of how well one knows e.g. a constant of nature
- engineer: tolerance = maximum possible deviation
- physicist: many different conventions...
  - ➜ standard deviation σ
  - ➜ 3-σ uncertainties
  - ➜ confidence level intervals containing the true value...
    - in a certain fraction of experiments (frequentist)
    - with a certain probability (bayesian)

➜ ask the professionals...

WG 1 (JCGM 100:2008, Recommendation INC-1 (1980))

➜ *Expression of experimental uncertainties*

1 The uncertainty in the result of a measurement generally consists of several components which may be grouped into two categories according to the way in which their numerical value is estimated:

A those which are evaluated by statistical methods,
B those which are evaluated by other means.

There is not always a simple correspondence between the classification into categories A or B and the previously used classification into“random” and “systematic” uncertainties. The term “systematic uncertainty” can be misleading and should be avoided. Any detailed report of the uncertainty should consist of a complete list of the components, specifying for each the method used to obtain its numerical value.

2 The components in category A are characterized by the estimated variances s_i^2 (or the estimated "standard deviations" s_i) and the number of degrees of freedom ν_i. Where appropriate, the covariances should be given.

3 The components in category B should be characterized by quantities u_j^2, which may be considered as approximations to the corresponding variances, the existence of which is assumed. The quantities u_j^2 may be treated like variances and the quantities u_j like standard deviations. Where appropriate, the covariances should be treated in a similar way.

4 The combined uncertainty should be characterized by the numerical value obtained by applying the usual method for the combination of variances. The combined uncertainty and its components should be expressed in the form of "standard deviations".

5 If, for particular applications, it is necessary to multiply the combined uncertainty by a factor to obtain an overall uncertainty, the multiplying factor used must always be stated.

(end of quote)

Discussion

➜ *using standard deviations to define uncertainties:*

- well defined simple procedures how to handle them
  - ➜ when propagating uncertainties into derived variables
  - ➜ for the combination of independent measurements
- rigorous limits on probability contents in the tails
- asymptotically gaussian behaviour
- no (little) danger of mis-interpretation
- profit from simple analytical properties of variances (linear functional)

➜ *using confidence level intervals to define uncertainties:*

- conservation of probability for monotonic transformations
- difficult to combine – requires likelihood function

❖ focus first on variances/standard deviations!

➜ *probability content in the tails of a distribution*

Take any PDF $f(x)$, function $w(x) \geq 0$ and x-region with $w(x) \geq C$:

$$\langle w \rangle = \int_{-\infty}^{\infty} dx\, f(x)\, w(x) \geq \int_{w(x) \geq C} dx\, f(x) w(x) \geq C \int_{w(x) \geq C} dx\, f(x)$$

$$\text{result:} \qquad p(x \text{ with } w(x) \geq C) \;\leq\; \frac{\langle w \rangle}{C}\,.$$

special choices: $w(x) = (x - \langle x \rangle)^2$ and $C = k^2\sigma^2$:

$$p_k \equiv p\left(x \text{ with } (x - \langle x \rangle)^2 > k^2\sigma^2\right) \leq \frac{1}{k^2}$$

- the probability beyond $\pm k\,\sigma$ around $\langle x \rangle$ is at most $1/k^2$
- gaussian PDF: $\{p_1, p_2, p_3\} \approx \{0.317, 0.0555, 0.0027\}$

➜ *definitions:*

- $\vec{x}$: vector of observed quantities
- $\langle\vec{x}\rangle$: expectation values of $\vec{x}$ - assumed to be the true values
- $C(x)$: covariance matrix of $\vec{x}$ - assumed to be known
- $\vec{y} = \vec{g}(\vec{x})$: vector of derived quantities
- $\vec{y}^{\,\mathrm{true}} = \vec{g}(\langle\vec{x}\rangle)$: true vector of derived quantities
- $C(y)$: covariance matrix of $\vec{y}$ - to be determined

❖ study properties of the transition $\vec{x} \rightarrow \vec{y}$

➜ expectation values

➜ uncertainties

➜ *the expectation value of $\vec{y}$ is biased: $\langle \vec{y} \rangle \neq \vec{y}^{\,\mathrm{true}}$*

Taylor expansion for a single component around $\langle x \rangle$ shows

$$y_k = g_k(\langle \vec{x} \rangle) + \sum_i \frac{\partial g_k(\langle \vec{x} \rangle)}{\partial x_i}(x_i - \langle x_i \rangle) + \frac{1}{2}\sum_{i,j} \frac{\partial^2 g_k(\langle \vec{x} \rangle)}{\partial x_i \partial x_j}(x_i - \langle x_i \rangle)(x_j - \langle x_j \rangle) + \ldots$$

and taking the expectation value yields:

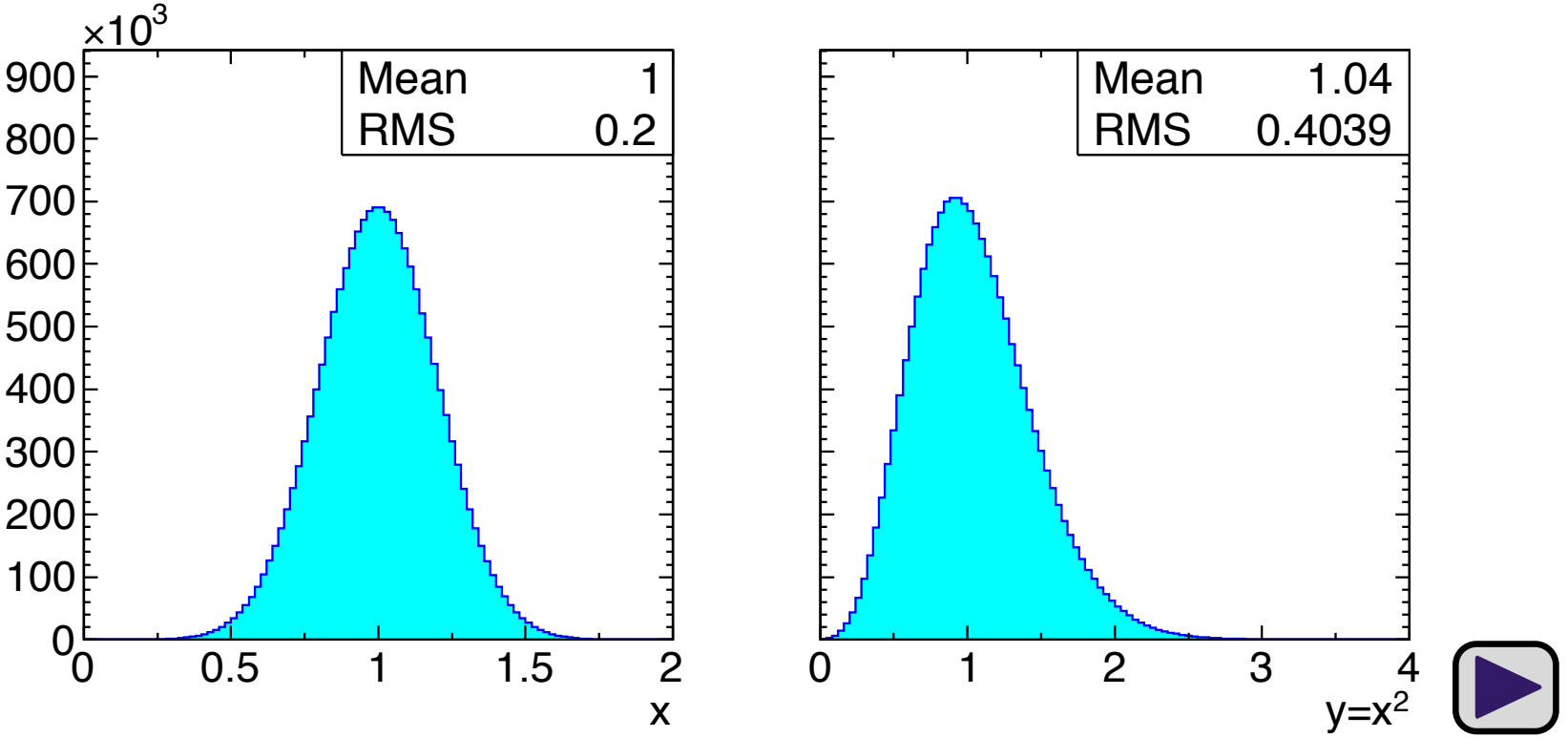
$$\langle y_k \rangle = y_k^{\mathrm{true}} + \frac{1}{2}\sum_{i,j} \frac{\partial^2 g_k(\langle \vec{x} \rangle)}{\partial x_i \partial x_j} C_{ij}(x) + \ldots$$

❖ discussion

- in many cases the bias is small and can be neglected
- the leading order correction in principle is known
- don't average biased estimates of $\vec{y}$ - average the unbiased $\vec{x}$

Numerical example

➜ *transformation of a gaussian distributed* $x \rightarrow y = x^n$



- small non-linearities or small σ are uncritical
- biases are usually small compared to standard deviations
- bias correction is needed when averaging transformed values

➜ *leading order treatment in n dimensions*

$$y_k \approx g_k(\langle\vec{x}\rangle) + \sum_{i=1}^{n} \frac{\partial g_k(\langle\vec{x}\rangle)}{\partial x_i}(x_i - \langle x_i\rangle) \qquad \text{expansion around } \langle\vec{x}\rangle$$

$$\approx g_k(\langle\vec{x}\rangle) + \sum_{i=1}^{n} \frac{\partial g_k(\vec{x})}{\partial x_i}(x_i - \langle x_i\rangle) \qquad \text{derivatives taken at } \vec{x}$$

➜ substitute $\vec{g}(\langle\vec{x}\rangle) = \vec{y}^{\,\text{true}}$

➜ covariance matrix estimate for bias corrected(!) transformed values

$$C_{kl}(y) = \left\langle (y_k - y_k^{\text{true}})(y_l - y_l^{\text{true}}) \right\rangle$$

$$= \sum_{i,j=1}^{n} \frac{\partial g_k}{\partial x_i}\frac{\partial g_l}{\partial x_j} \langle (x_i - \langle x_i\rangle)(x_j - \langle x_j\rangle)\rangle = \sum_{i,j=1}^{n} \frac{\partial g_k}{\partial x_i}\frac{\partial g_l}{\partial x_j} C_{ij}(x)$$

Matrix notation

For the transformation $\vec{y} = \vec{g}(\vec{x})$ with Jacobian

$$M(x) \qquad \text{and matrix elements} \qquad M_{ij} = \frac{\partial g_i}{\partial x_j}$$

the transformed covariance matrix becomes

$$C_{kl}(y) = \sum_{i,j} M_{ki} M_{lj} C_{ij}(x) \quad \text{or} \quad C(y) = M(x) \cdot C(x) \cdot M^T(x)$$

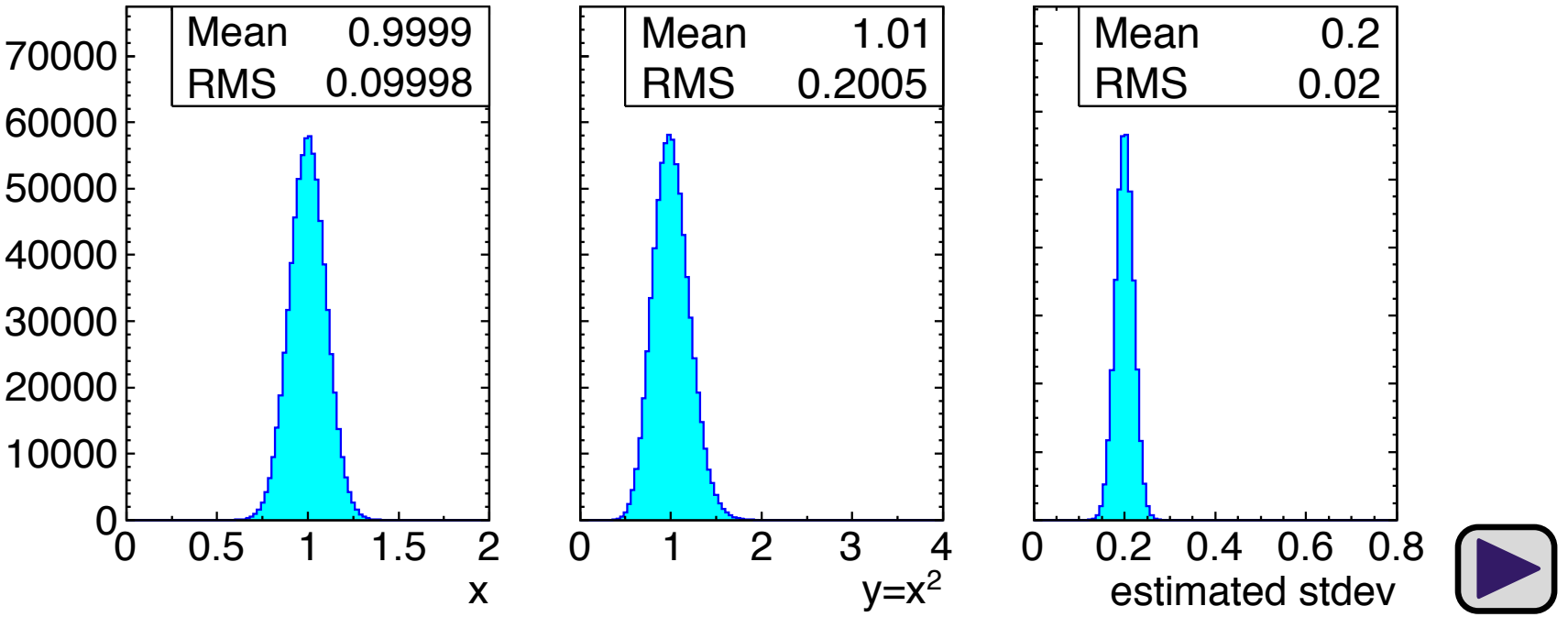
- the argument to M defines that the derivatives are w.r.t. x
- for invertible $M(x)$ no information is lost in the transformation
- chaining transformations leads to

$$\vec{y} = \vec{h}(\vec{g}(\vec{x})) \quad \text{and} \quad M_{ij} = \sum_{k=1}^{n} \frac{\partial h_i}{\partial g_k} \frac{\partial g_k}{\partial x_j} \quad \text{or} \quad M = M(g) \cdot M(x)\,.$$

➜ identical results if doing transformations in one or many steps

Numerical studies (i)

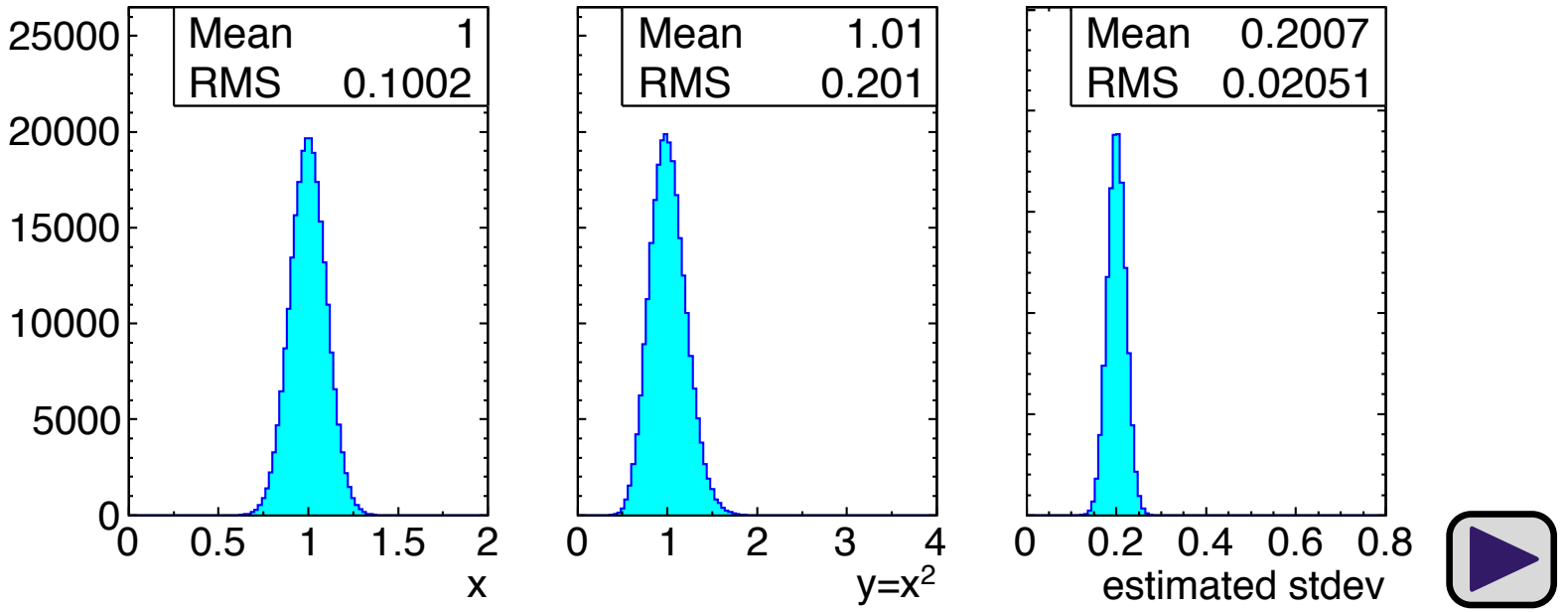
➜ *estimated and exact standard deviations for* $x \to y = x^n$



- average error estimates are OK
- actual values scatter proportional to relative errors of x
  - ❖ uncertainties are inherently uncertain

➜ *toy-MC and exact standard deviations for* $x \rightarrow y = x^n$

Fluctuate every measured value x by its known variance and estimate the standard deviation of y from the transformed x-values.



- similar behaviour as analytical results (slightly larger scatter)
- easy to implement as no derivatives are required
- small sensitivity to PDF of fluctuations

Summary: dealing with standard deviations

➜ *linear error propagation:*

- requires only covariance matrix of the input
- exact for linear and approximate for non-linear transformations
- higher order corrections need higher order moments of the input
  - ➜ behaviour like an asymptotic series ➜ numerically diverging
  - ➜ no improvement w.r.t. leading order treatment
- consistent when chaining transformations
- equivalent quality as error propagation via toy-MC
- the non-constant Jacobian of non-linear transformations induces errors for transformed variances - even for known input variances
- non-linear transformations induce bias
  - ➜ leading order correction of transformed values is recommended
  - ➜ no bias correction is needed for transformed covariance matrices

➜ *alternative ways to quantify uncertainties*

- no longer distribution-free – the underlying PDFs need to be known
- propagation of confidence level intervals is easy
- combination of uncertainties not well defined
  - ➜ common practice:

$$a = 42 \pm^{8}_{3} \pm^{6}_{4} = 42 \pm^{10}_{5}$$

  - ➜ little or no theoretical backing
  - ➜ implies the concept of asymmetric variance
  - ➜ implies that confidence level intervals behave like variances
- different concepts in bayesian and frequentist frameworks

a simple case study ➜

➜ *setting the scene:*

A counting experiment has observed n events. The experiment recorded independent random processes that occur with a constant probability per time interval, such as e.g. radiocative decays. It thus is known that n is a poissonian distributed random variable, i.e. the probability P_n to observe n events is:

$$P_n = P(n;\mu) = e^{-\mu}\,\frac{\mu^n}{n!}$$

➜ *question:*

What can be inferred about the expectation value μ?

Example: $n = 2$

➜ *quick check of a few hypotheses . . .*

- ➜ $P(2; \mu = 0.1) \approx 0.0045$
- ➜ $P(2; \mu = 1.0) \approx 0.1839$
- ➜ $P(2; \mu = 10.) \approx 0.0023$

- in principle any value for μ is possible
- a value $\mu = O(1)$ seems more plausible

❖ try to be quantitative about a certain range of μ

- discuss
  - ➜ the bayesian approach
  - ➜ the frequentist approach

➜ *treat μ as a random variable*

- formally possible even if μ has a well defined true physical value
- interpret the PDF of μ as encoding the knowledge about μ
- use Bayes' theorem to improve the knowledge by the measurement:

$$P(\mu|n)\,P(n) = P(n|\mu)\,P(\mu)$$

  - ➜ $P(\mu)$: prior PDF of μ - to be defined
  - ➜ $P(n|\mu)$: Likelihood function
  - ➜ $P(n)$: probability for n, unknown constant
  - ➜ $P(\mu|n)$: posterior PDF for μ after the measurement

❖ it follows

$$P(\mu|n) \propto P(n|\mu)\,P(\mu) \quad \text{and thus} \quad P(\mu|n) = \frac{P(n|\mu)\,P(\mu)}{\int d\mu\; P(n|\mu)\,P(\mu)}$$

➜ *choice of prior distribution*

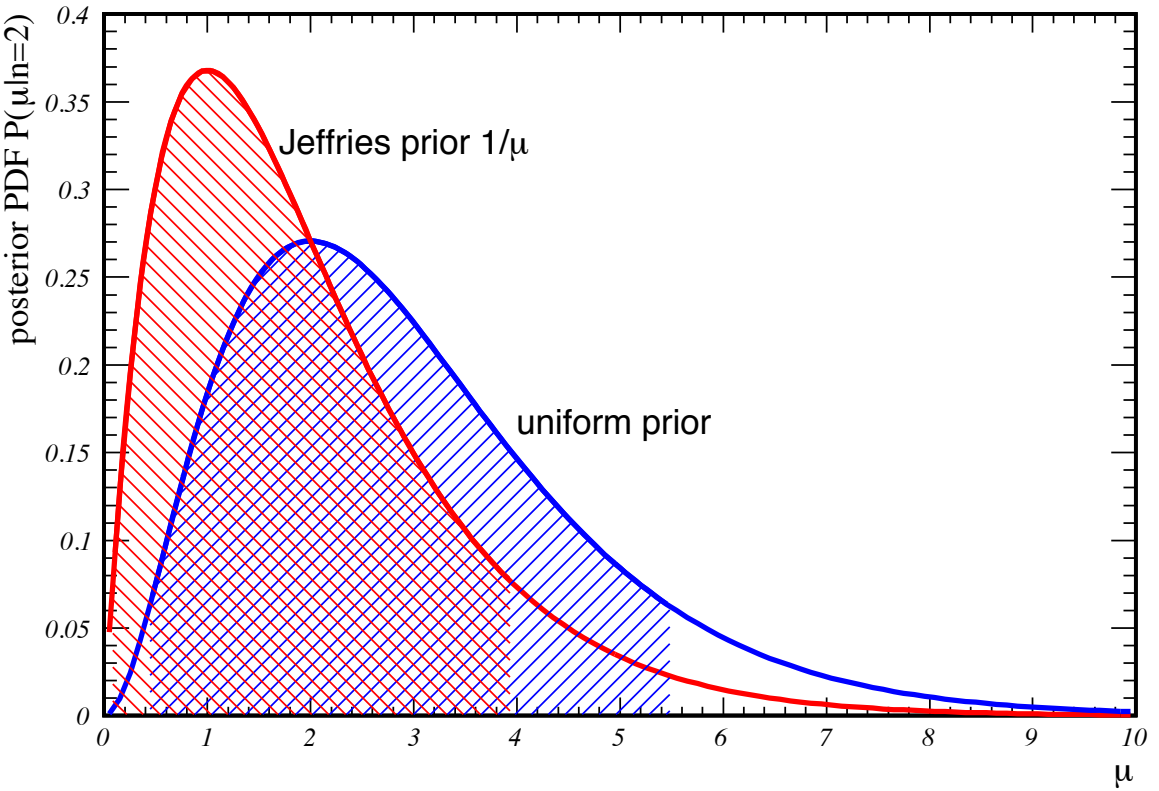
$$P(\mu) = \mu^k$$

- ad hoc - but allows to test sensitivity to prior, special cases:
- $k = 0$: equal probability for all possible values
- $k = -1$ Jeffries prior: invariance w.r.t scale-transformations $\mu \to \alpha\,\mu$

$$P(\mu|n) = \frac{e^{-\mu}\,\mu^{n+k}}{\int_0^\infty d\mu\; e^{-\mu}\,\mu^{n+k}} = e^{-\mu}\,\frac{\mu^{n+k}}{(n+k)!}$$

results ➜

Bayesian 90% confidence level intervals



- 90% confidence intervals are regions with 90% probability content
  - ➜ many possibilities - usually take the smallest interval
- most probable values and confidence intervals sensitive to prior

Discussion

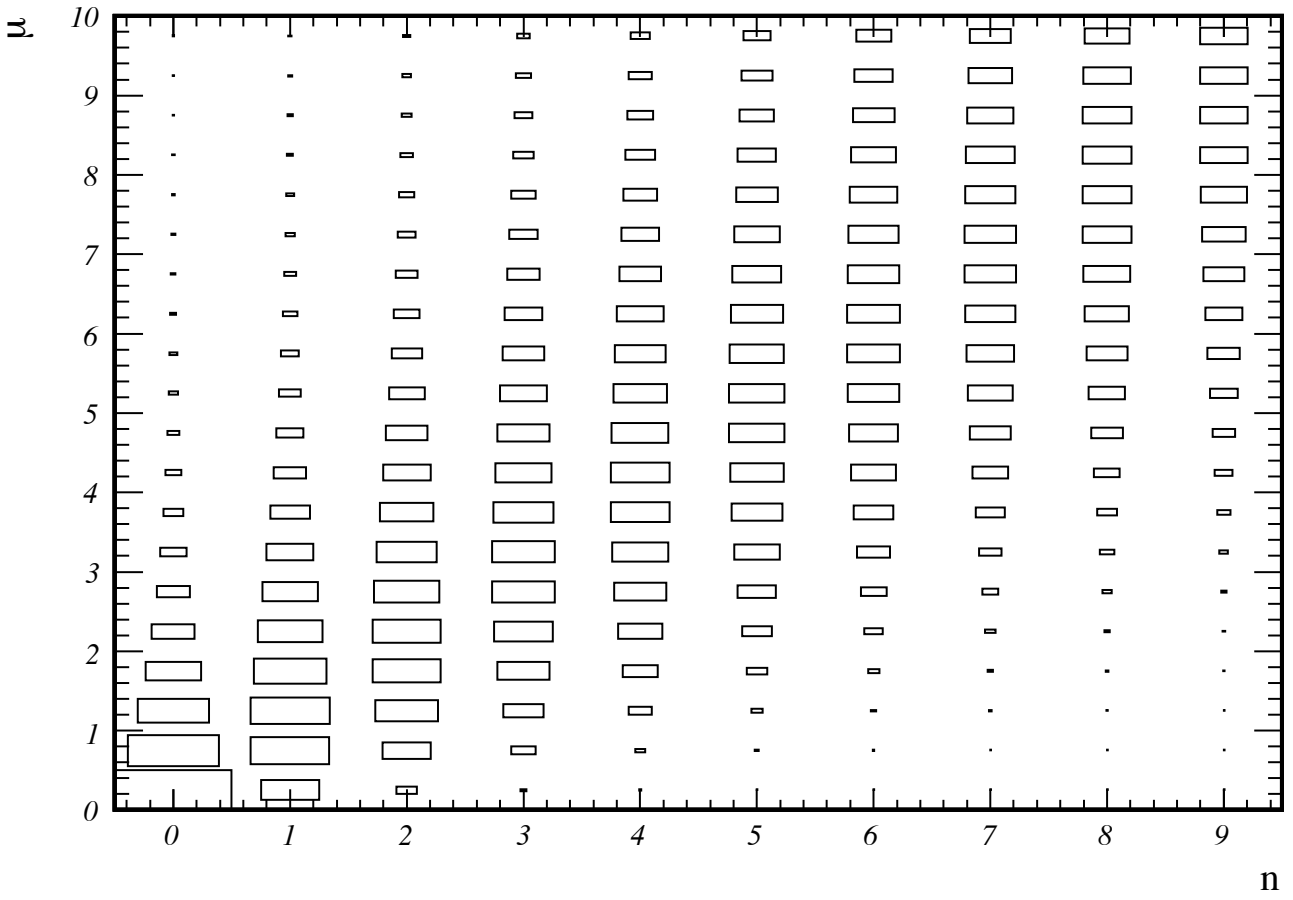
- bayesian approach formalizes gain of knowledge by measurement
  - ➜ posterior of first measurement can be prior of second, etc.

$$P(\mu|n_2, n_1) \propto P(n_2|\mu)P(n_1|\mu)P(\mu)$$
$$= P(n_2, n_1|\mu)P(\mu)$$
$$= P(n_2|\mu)P_1(\mu) \quad \text{with} \quad P_1(\mu) = P(n_1|\mu)P(\mu)$$

- consistent if a non-uniform prior (e.g. Jeffries') is used only once
  - ➜ avoid non-uniform priors for single measurements
  - ➜ if needed, use a non-uniform prior once when combining results
  - ➜ possible if likelihood functions are published
- caveat: uniformity depends on the definition of the parameter
  - ➜ example: uniform in μ is non-uniform in $\sqrt{\mu}$

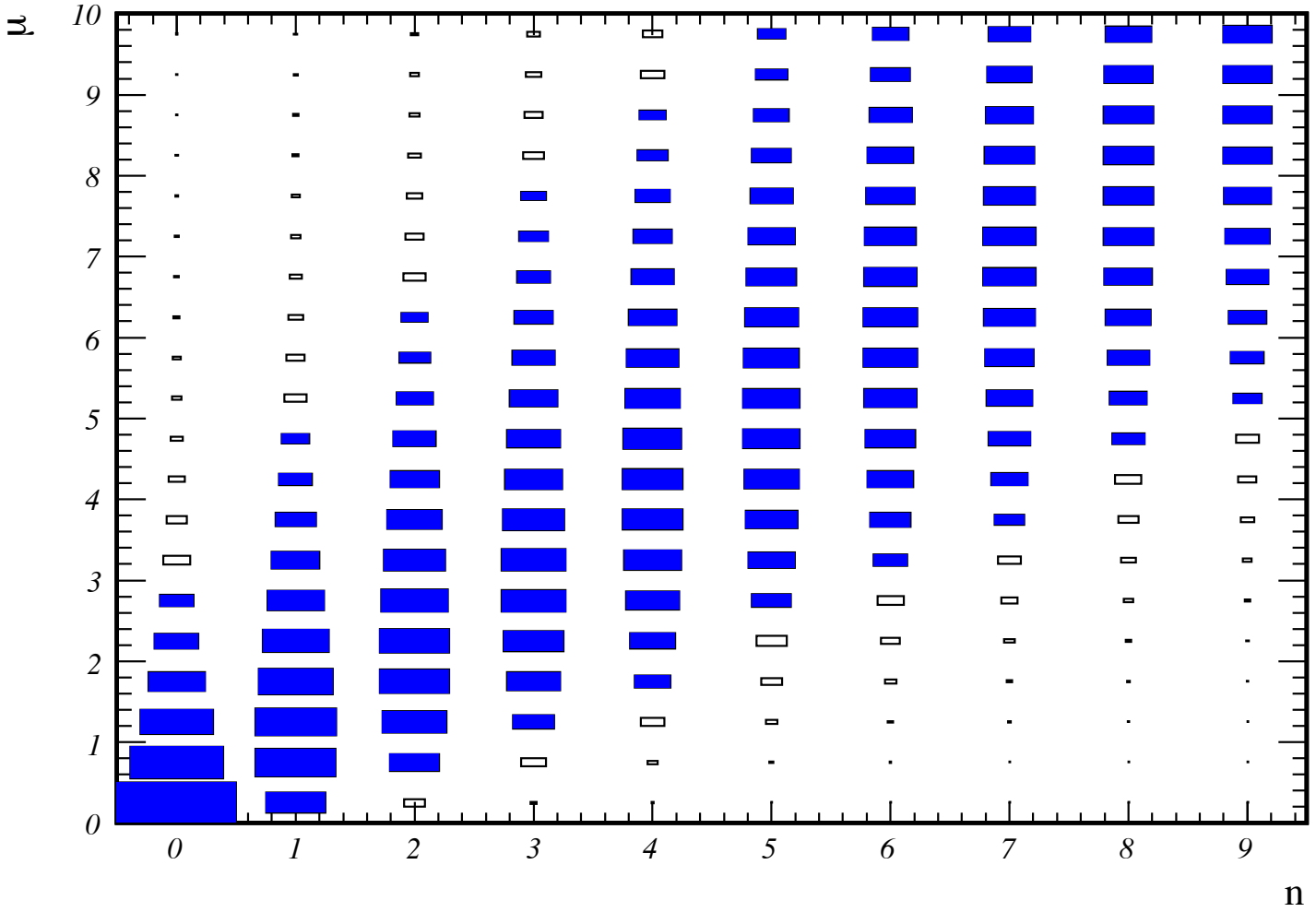
➜ *Likelihood-function-only based "Neyman construction"*

- start from table of probabilities for any observation and any vaue μ



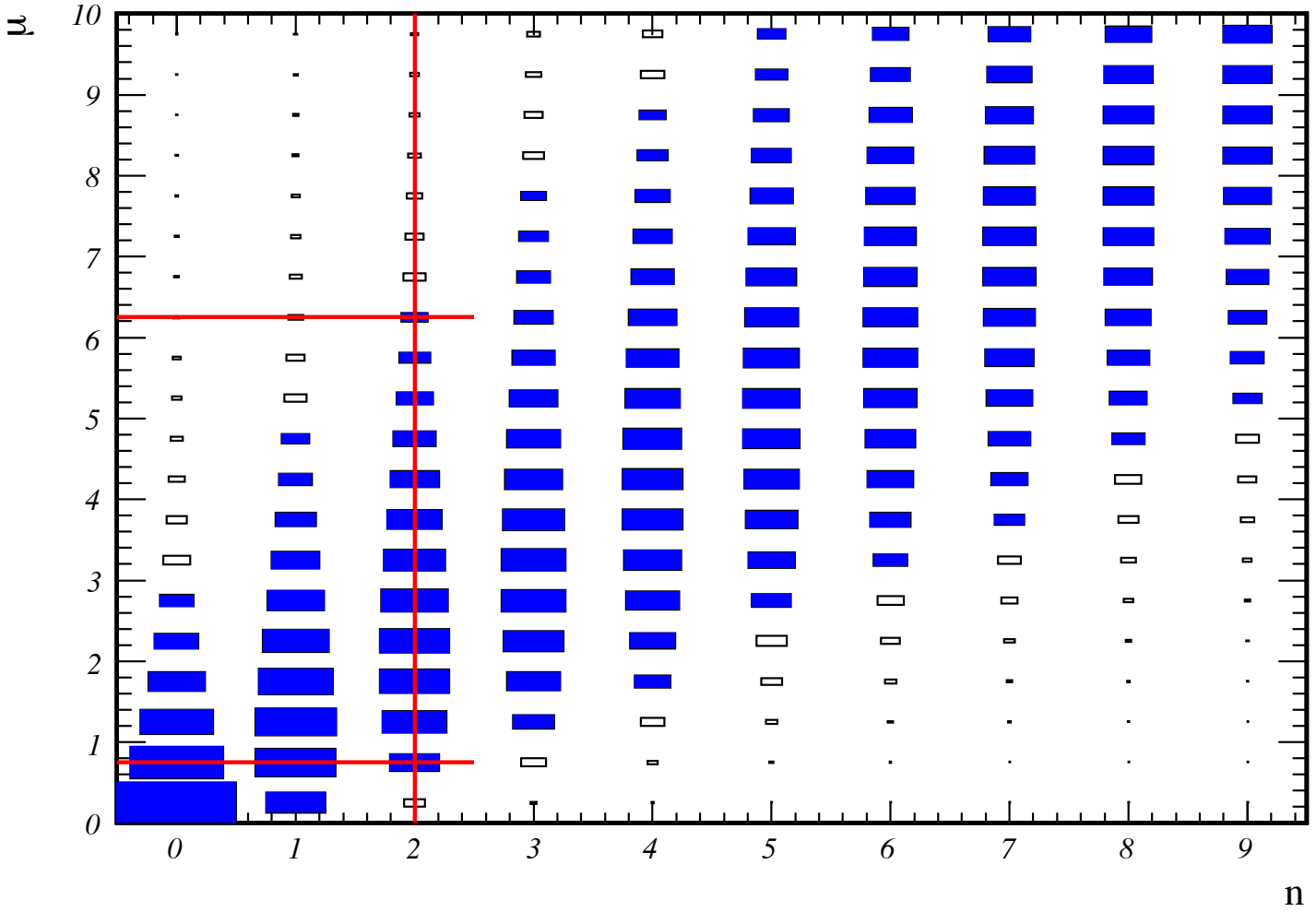
The Neyman construction (i)

- determine the shortest $\geq 90\%$ horizontal range for each μ



The Neyman construction (ii)

- given n, take the range of μ with n in the $\geq 90\%$ probability range



Properties of the Neyman construction

- a fixed interval for μ is assigned to every measurement n
- every interval contains the true value with 90% probability
  - ➜ false from the frequentist point of view – the true value μ is either inside or outside; there are no probabilities for this.
- from an ensemble of measurements (at least) 90% of the confidence level intervals are expected to contain the true value
  - ➜ true – for any true μ, different measurements will find different values n and thus will quote different intervals. Take for example $\mu = 4.25$. It is contained in the intervals of $n = 1, \ldots, 7$, and by construction, (at least) 90% of the measurements are in that range. Analogous reasoning holds for all μ.
- the interval contains no information about preferred values!

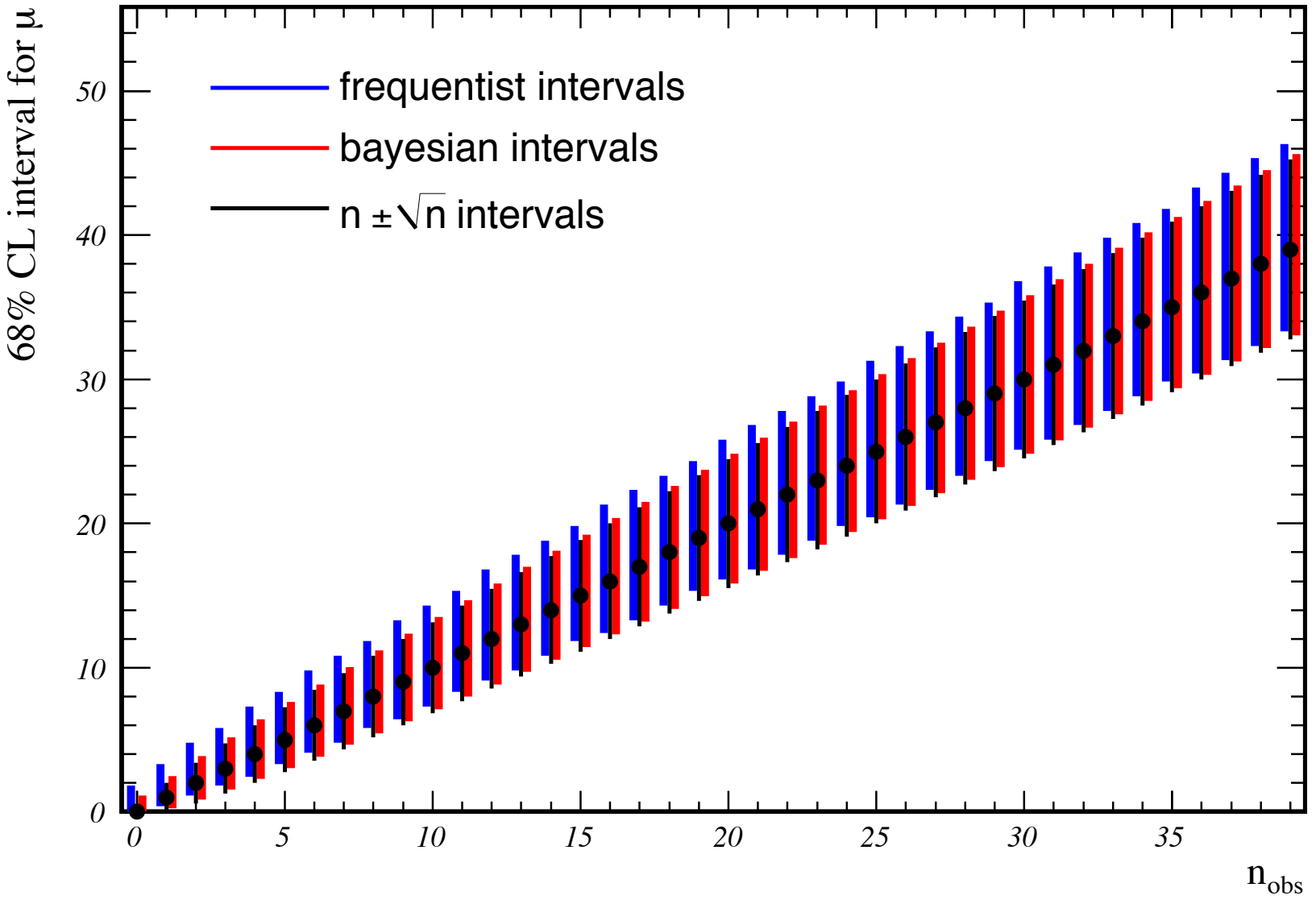
➜ *some common themes. . .*

- bayesian and frequentist methods define regions $[\mu_l(n), \mu_h(n)]$
- for each observation n there is a well defined interval
- another common region for 68% confidence level is $n \pm \sqrt{n}$

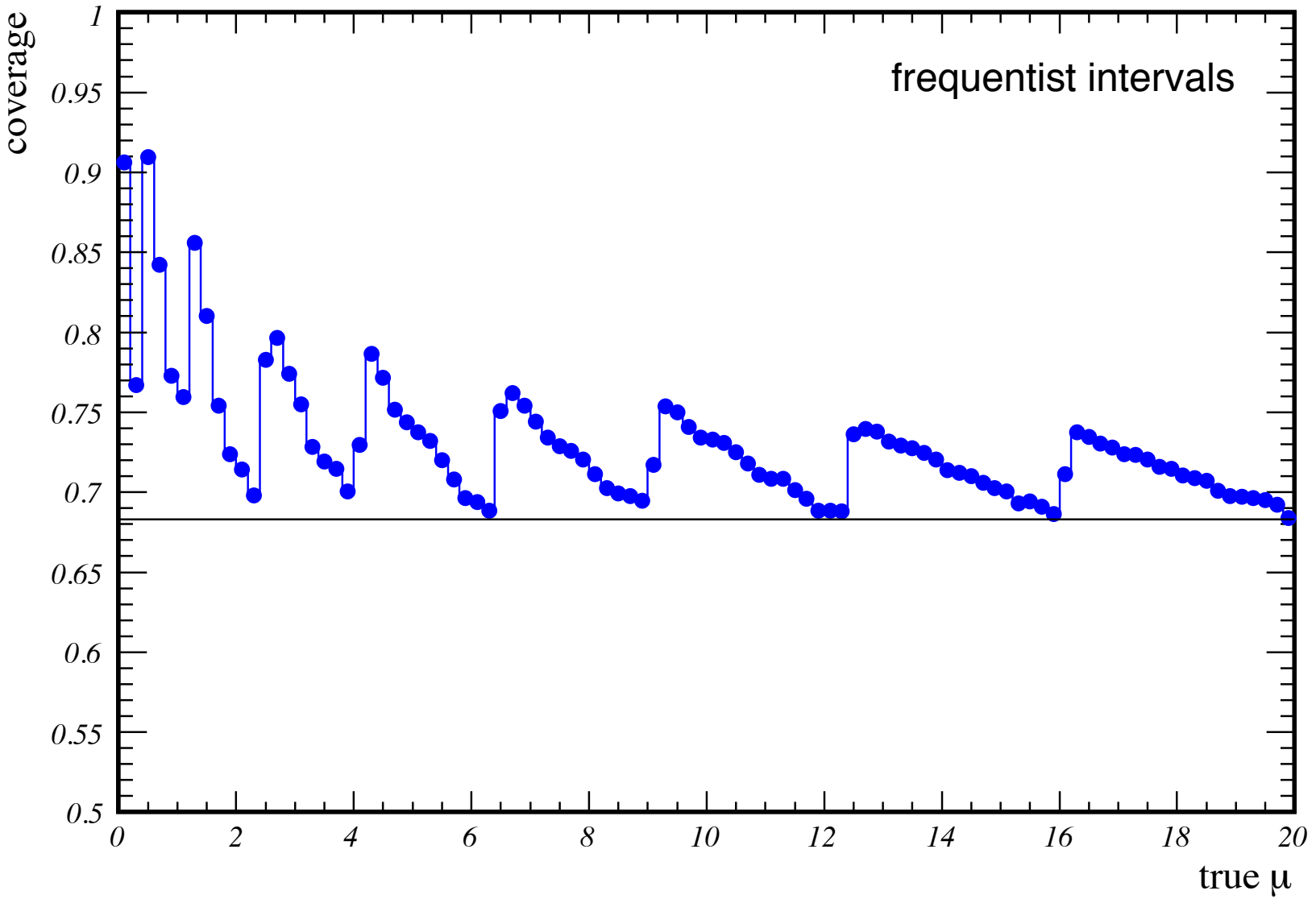
➜ *further studies. . .*

- compare the intervals defined by the different schemes
- MC check which fraction of intervals contain the true value
  - ➜ do the check as a function of the unknown true μ
  - ➜ check that the frequentists intervals have coverage
  - ➜ calculate coverage also for bayesian intervals
    - even if bayesians do not care about coverage . . .

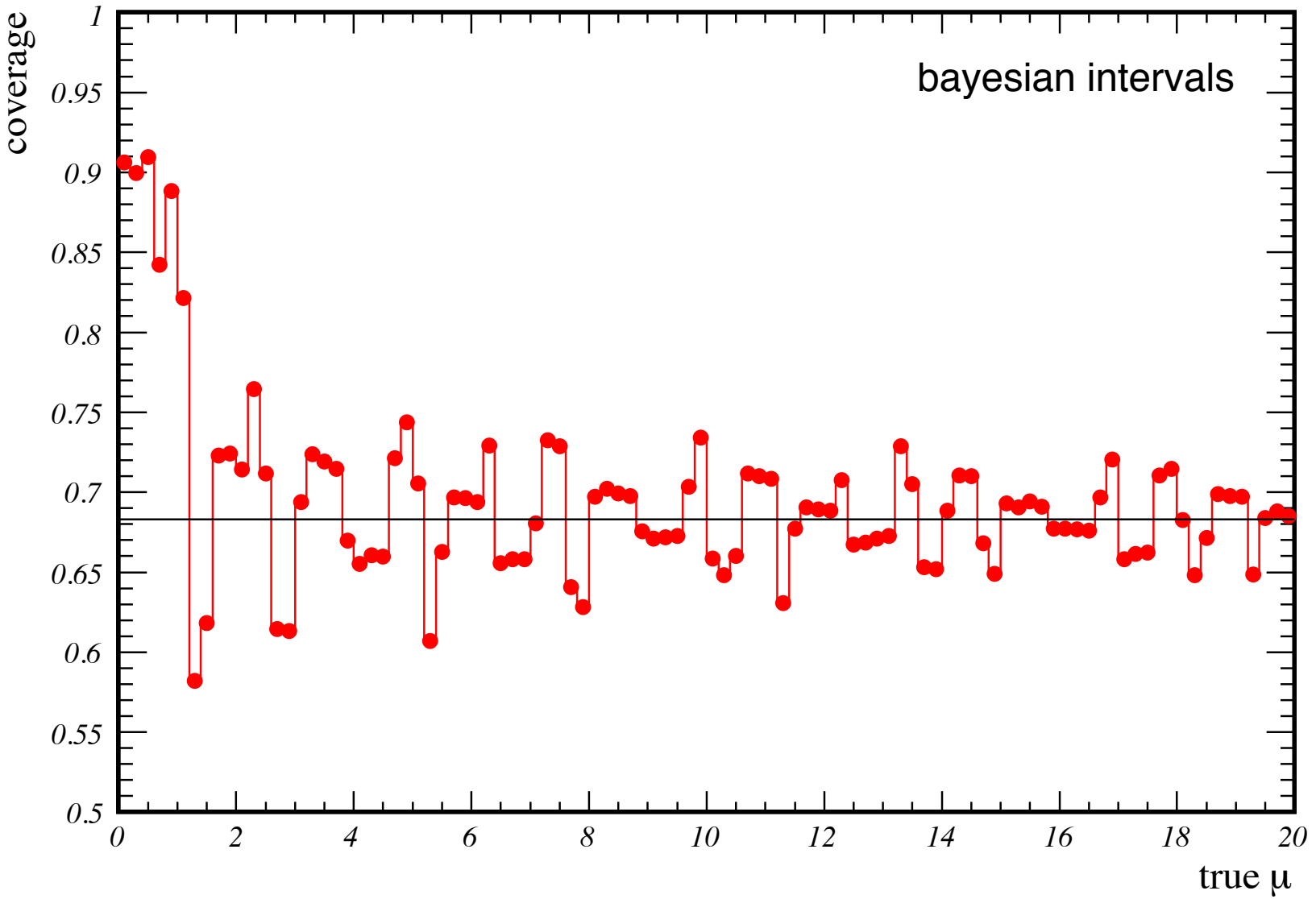
68.3% confidence level intervals



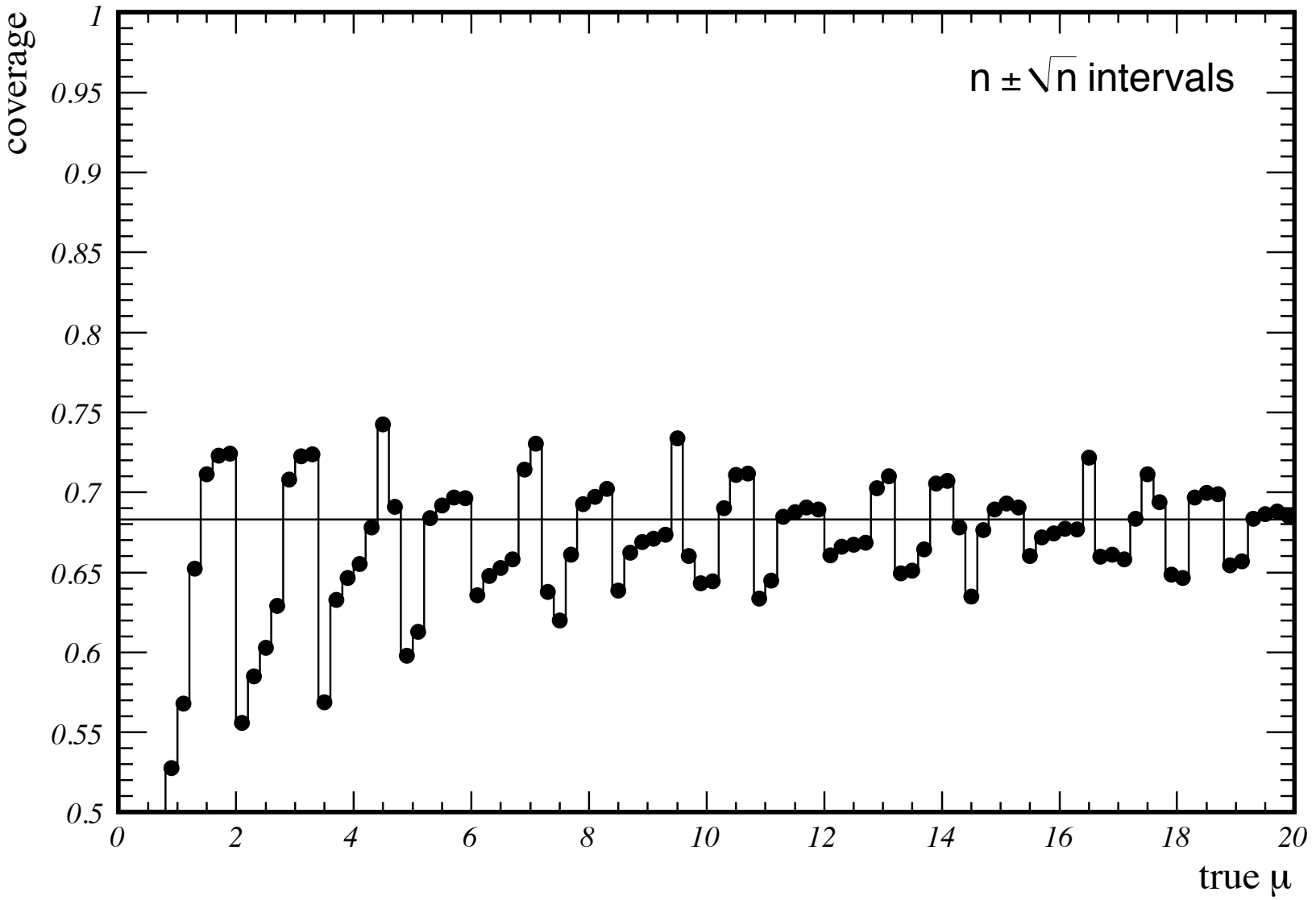
Coverage of 68.3% confidence level intervals



Coverage of 68.3% confidence level intervals



Coverage of 68.3% confidence level intervals



Concluding remarks

- bayesians makes statements about the theory
  - ➜ "The true value μ is with 90% probability inside the 90% confidence level interval"
- frequentists makes statements about the data
  - ➜ "90% of the 90% confidence level intervals (which are a function of the data) are expected to contain the true value μ"
  - ➜ these confidence level intervals have "coverage"
  - ➜ for continuous PDFs exact coverage can be obtained
  - ➜ discrete probabilities are chosen to have over-coverage
- bayesians & frequentists base CL-intervals on the likelihood function
- confidence level intervals from maximum likelihood or least squares fits based on $\Delta\chi^2$ or $\Delta \ln L$ are exact only for gaussian PDFs. In most cases they don't have coverage.
- treating confidence level intervals like variances is questionable

➜ *extract physics parameters from a set of measurements*

❖ properties which are assumed to be satisfied:

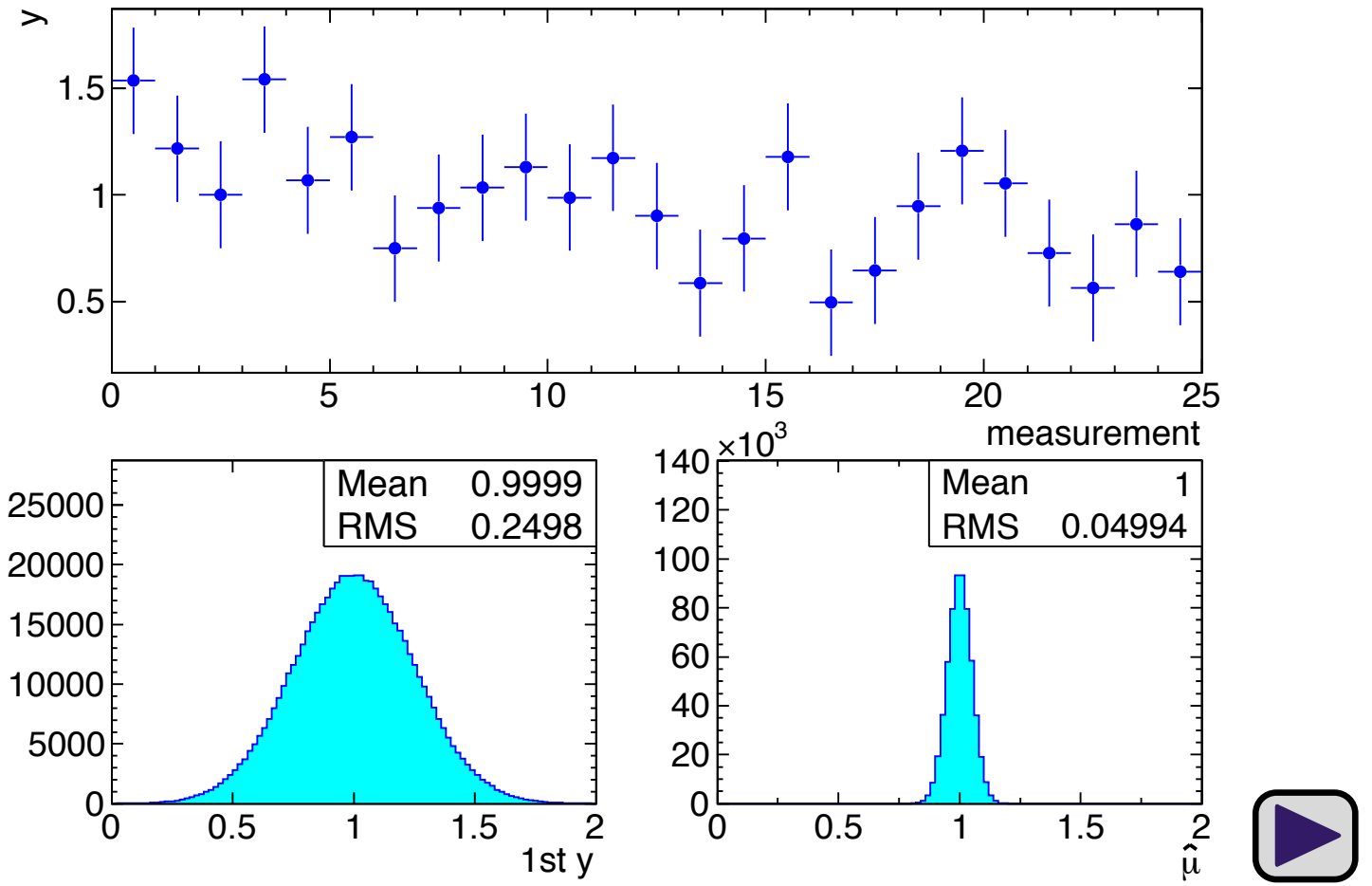
- individual measurements fluctuate with known variance
- individual measurements are unbiased

➜ *measurements of the same physical quantity*

- scenario
  - ➜ n measurements y_i with $i = 1, 2, \ldots, n$
  - ➜ all measurements fluctuate around an unknown true value μ
  - ➜ the measurements have standard deviations σ_i
- each measurement is an estimate for μ with uncertainty σ_i
- task: combine the measurements for a better estimate of μ

➜ try the arithmetic average $\hat{\mu} = \frac{1}{n} \sum_{i=1}^{n} y_i$

Numerical simulation



- big improvements if all variances are the same
- less/no improvement w.r.t. best measurement for different variances

➜ *modification of the arithmetic average*

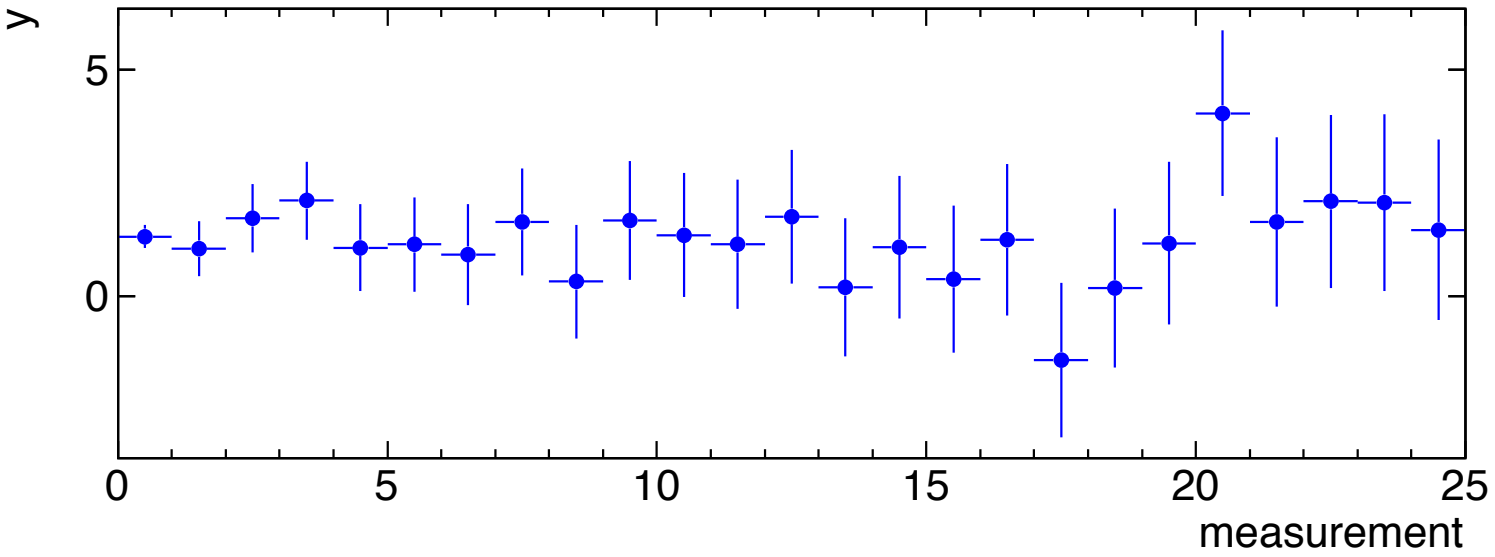
$$\hat{\mu} = \sum_{i=1}^{n} w_i y_i \qquad \text{with} \qquad \sum_{i=1}^{n} w_i = 1$$

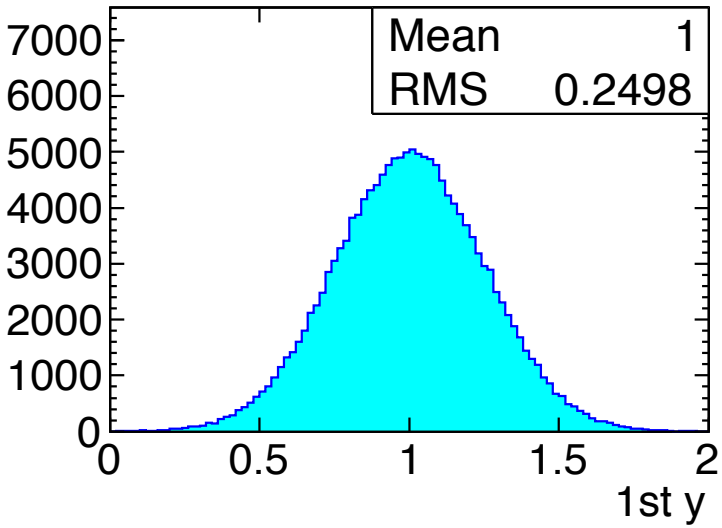
- consistent results for arbitrary weights: $\hat{\mu} = \mu$ if $y_i = \mu$
- try to find weights which minimize the variance of $\hat{\mu}$

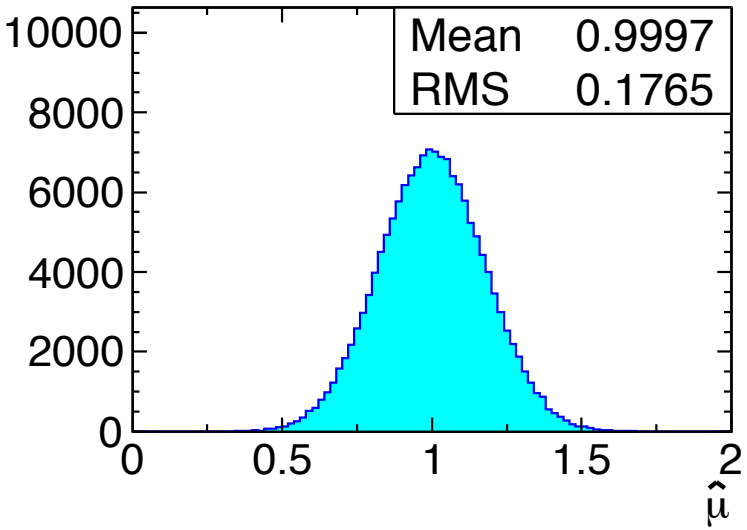
$$\sigma^2(\hat{\mu}) = \sum_{i=1}^{n} w_i^2 \sigma_i^2 \stackrel{!}{=} \min$$

- constrained minimization problem
- minimum for $w_i \propto 1/\sigma_i^2$
- recovers unweighted average if all σ_i are the same

Numerical simulation







➜ *use case: straight line fit*

Consider uncorrelated measurements $y_i, i = 1, \ldots, n$ with known variances σ_i^2, recorded for certain values x_i of a control paramater x. The expectation value of the measurements is $\langle y_i \rangle = a_0 + a_1 x_i$, where the parameters a_0 and a_1 are not known.

➜ *wanted: a method to find estimates $\hat{a}_0$ and $\hat{a}_1$ for a_0 and a_1*

❖ discussion

- control parameters x_i are known
- the measurements y_i are unbiased
- variances σ_i^2 are known
- exact shape of PDFs describing the fluctuations of the y_i is irrelevant
  - ➜ any PDF with variance σ_i^2 would do
  - ➜ different measurements can fluctuate with different PDFs

➜ *the case of two measurements*

$$\langle y_1 \rangle = a_0 + a_1\, x_1 \quad \text{and} \quad y_1 = \langle y_1 \rangle + r_1$$
$$\langle y_2 \rangle = a_0 + a_1\, x_2 \quad \text{and} \quad y_2 = \langle y_2 \rangle + r_2$$

- system of linear equations relating $\langle y_i \rangle$ and x_i
- measurements y_i have random deviation r_i from $\langle y_i \rangle$
- unbiasedness of y_i implies $\langle r_i \rangle = 0$
- estimate a_0 and a_1 by assuming $r_i = 0$, i.e. make the ansatz:

$$y_1 = \hat{a}_0 + \hat{a}_1\, x_1$$
$$y_2 = \hat{a}_0 + \hat{a}_1\, x_2$$

❖ result:

$$\hat{a}_0 = y_1 - \hat{a}_1 x_1 = \frac{x_2}{x_2 - x_1}\, y_1 - \frac{x_1}{x_2 - x_1}\, y_2$$
$$\hat{a}_1 = \frac{y_2 - y_1}{x_2 - x_1} = -\frac{1}{x_2 - x_1}\, y_1 + \frac{1}{x_2 - x_1}\, y_2$$

➜ *does the estimate make sense?*

- parameter estimates are linear combinations of the measurements
- parameter estimates are random variables
- parameter estimates fluctuate with the measurements
- check the expectation values ...

$$\langle \hat{a}_0 \rangle = \left\langle \frac{1}{x_2 - x_1}(x_2 y_1 - x_1 y_2) \right\rangle = \frac{1}{x_2 - x_1}(x_2 \langle y_1 \rangle - x_1 \langle y_2 \rangle) = a_0$$

$$\langle \hat{a}_1 \rangle = \left\langle \frac{1}{x_2 - x_1}(-y_1 + y_2) \right\rangle = \frac{1}{x_2 - x_1}(-\langle y_1 \rangle + \langle y_2 \rangle) = a_1$$

❖ conclusion:

➜ the estimates for the unknown parameters are unbiased

➜ the parameter errors can be determined by error propagation

yes, the parameter estimates make sense!

➜ *the case of $n > 2$ measurements*

Take the lessons learnt from the case $n = 2$ and try to estimate the unknown parameters by a linear combination of the measurements.

$$\hat{a}_0 = \sum_{i=1}^{n} p_i \, y_i \quad \text{and} \quad \hat{a}_1 = \sum_{i=1}^{n} q_i \, y_i$$

- this is a convenient ansatz, not derived from any “first principles”
- it is not the only possible generalization of the case $n = 2$
- nor will it give the best possible estimates for a_0 and a_1
- but it is simple and robust, requiring only minimal input
- and turns out to be surprisingly powerful . . .

➜ determine parameters p_i and q_i . . .

Optimizing the parameter estimates

➜ *exploit the freedom of the linear ansatz to. . .*

- make sure that the estimates are unbiased
- and that the estimates are as accurate as possible

❖ condition for unbiased estimates:

$$\langle \hat{a}_0 \rangle = \sum_{i=1}^{n} p_i \langle y_i \rangle = \sum_{i=1}^{n} p_i (a_0 + a_1 x_i) = a_0 \sum_{i=1}^{n} p_i + a_1 \sum_{i=1}^{n} p_i x_i \stackrel{!}{=} a_0$$

$$\langle \hat{a}_1 \rangle = \sum_{i=1}^{n} q_i \langle y_i \rangle = \sum_{i=1}^{n} q_i (a_0 + a_1 x_i) = a_0 \sum_{i=1}^{n} q_i + a_1 \sum_{i=1}^{n} q_i x_i \stackrel{!}{=} a_1$$

one obtains 4 conditions:

$$\sum_{i=1}^{n} p_i = 1 \qquad \sum_{i=1}^{n} q_i = 0 \qquad \sum_{i=1}^{n} p_i x_i = 0 \qquad \sum_{i=1}^{n} q_i x_i = 1$$

- only 4 constraints for $2n$ parameters
- easy to satisfy both for p_i and q_i
  - ➜ start from a set of random numbers e.g. for p_i
  - ➜ subtract a constant such that the “0-constraint” is satisfied
  - ➜ scale the numbers such that the “1-constraint” is satisfied
- additional criterion needed to fix the coefficients
- require minimal variance for the parameter estimates
  - ➜ constrained minimization problem

❖ variance of parameter estimates from error propagation:

$$\sigma^2(\hat{a}_0) = \sum_{i=1}^{n} \left(\frac{\partial \hat{a}_0}{\partial y_i} \right)^2 \sigma_i^2 = \sum_{i=1}^{n} p_i^2 \, \sigma_i^2 \quad \text{and} \quad \sigma^2(\hat{a}_1) = \sum_{i=1}^{n} q_i^2 \, \sigma_i^2$$

➜ result of a constrained minimization

➜ *a little algebra yields the textbook formulae for straight line fits:*

Expressed through the following sums

$$S_{\{1,x,xx,y,xy\}} = \sum_{i=1}^{n} \frac{\{1, x_i, x_i\, x_i, y_i, x_i\, y_i\}}{\sigma_i^2} \quad \text{and} \quad D = S_1 S_{xx} - S_x^2$$

the best-fit estimates are

$$\hat{a}_0 = \frac{1}{D}(S_{xx} S_y - S_x S_{xy}) \quad \text{and} \quad \hat{a}_1 = \frac{1}{D}(S_1 S_{xy} - S_x S_y)$$

and error propagation yields the covariance matrix elements

$$C_{kl}(\hat{a}) = \sum_{i=1}^{n} \frac{\partial \hat{a}_k}{\partial y_i} \frac{\partial \hat{a}_l}{\partial y_i} \sigma_i^2$$

$$C_{00}(\hat{a}) = \frac{S_1}{D} \quad , \quad C_{11}(\hat{a}) = \frac{S_{xx}}{D} \quad \text{and} \quad C_{01}(\hat{a}) = \frac{-S_x}{D}$$

➜ *re-write the solution derived before...*

$$\hat{a}_0 = \frac{1}{D}(S_{xx}S_y - S_x S_{xy}) \quad \text{and} \quad \hat{a}_1 = \frac{1}{D}(S_1 S_{xy} - S_x S_y)$$

to make the structure more evident:

$$\begin{pmatrix} \hat{a}_0 \\ \hat{a}_1 \end{pmatrix} = \frac{1}{D}\begin{pmatrix} S_{xx} & -S_x \\ -S_x & S_1 \end{pmatrix} \cdot \begin{pmatrix} S_y \\ S_{xy} \end{pmatrix} \quad \rightarrow \quad \begin{pmatrix} S_1 & S_x \\ S_x & S_{xx} \end{pmatrix} \cdot \begin{pmatrix} \hat{a}_0 \\ \hat{a}_1 \end{pmatrix} = \begin{pmatrix} S_y \\ S_{xy} \end{pmatrix}$$

or

$$S_1\ \hat{a}_0 + S_x\ \hat{a}_1 - S_y\ = 0$$
$$S_x\ \hat{a}_0 + S_{xx}\ \hat{a}_1 - S_{xy} = 0$$

Two equations which define the best fit parameters as the zero of a two-dimensional function, or equivalently, as a stationary point (e.g. minimum) of its primitive χ^2.

➜ *integration:*

$$\chi^2 = S_1\, a_0^2 + S_{xx}\, a_1^2 + 2S_x\, a_0\, a_1 - 2S_y\, a_0 - 2S_{xy}\, a_1 + K$$

- quadratic form with a free integration constant K
- asking $\chi^2_{\text{min}} = 0$ yields $K = \sum y_i^2/\sigma_i^2$
  - ➜ ideal fits have zero cost
  - ➜ allows to interpret the best-fit χ^2 as a quality of fit

Result:

$$\chi^2 = \sum_{i=1}^{n} \frac{(y_i - a_0 - a_1 x_i)^2}{\sigma_i^2}$$

$$= \sum_{i=1}^{n} \frac{(y_i - f_i(a_0, a_1))^2}{\sigma_i^2} \quad \text{with} \quad f_i(a_0, a_1) = a_0 + a_1 x_i$$

Discussion

- minimize data−model measured in units of standard deviations
- the derivation was for uncorrelated data points y_i
- noting $1/\sigma_i^2 = C_{ii}^{-1}(y)$, the general expression becomes

$$\chi^2 = \sum_{i,j=1}^{n} (y_i - f_i(a_0, a_1))\, C_{ij}^{-1}(y)\, (y_j - f_j(a_0, a_1))$$

➜ using matrix notation, with $C_y \equiv C(y)$,

$$\chi^2 = \vec{r}^{\,T}\; C_y^{-1}\; \vec{r} \qquad \text{with} \qquad \vec{r} = \vec{y} - \vec{f}(a_0, a_1)$$

- invariance under linear transformations L

$$\vec{r}\,' = L\,\vec{r} \quad , \quad C_y' = L\,C_y\,L^T \quad , \quad C_y'^{-1} = (L^T)^{-1}\,C_y^{-1}\,L^{-1}$$

$$\text{and thus} \qquad (\chi^2)' = \chi^2$$

➜ allows simplification by diagonalizing $C(y)$

➜ *(average) measurements are linear functions of parameters $\vec{a}$*

$$\chi^2 = (\vec{y} - M\,\vec{a})^T\ C_y^{-1}\,(\vec{y} - M\,\vec{a})$$
$$= \vec{y}^{\,T} C_y^{-1} \vec{y} - 2\vec{a}^{\,T}\left[M^T C_y^{-1}\vec{y}\right] + \vec{a}^{\,T}\left[M^T C_y^{-1} M\right]\vec{a}$$

The best fit parameters are linear functions of the measurements

$$\vec{a} = Q\ \vec{y} \qquad \text{with} \qquad Q = \left[M^T C_y^{-1} M\right]^{-1} M^T C_y^{-1}$$

with covariance matrix

$$C_a = Q\ C_y\ Q^T = \left[M^T C_y^{-1} M\right]^{-1} = \left(\frac{1}{2}\frac{\partial\chi^2}{\partial\vec{a}^2}\right)^{-1}.$$

- using C_y^{-1} in the χ^2 function gives the best fit (BLUE)
- other constant matrices are possible as well
- use $C_a = QC_yQ^T$ to get the correct covariance matrix

- unbiased parameter estimates (for any constant matrix C_y^{-1})

$$\langle \vec{y} \rangle = M \vec{a}_{\text{true}} \quad \rightarrow \quad \langle \vec{a} \rangle = \left[M^T C_y^{-1} M \right]^{-1} M^T C_y^{-1} \langle \vec{y} \rangle = \vec{a}_{\text{true}}$$

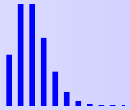
- minimum χ^2 value

$$\begin{aligned} \chi^2_{\text{min}} &= \vec{y}^T C_y^{-1} \vec{y} - \vec{a}^T \left[M^T C_y^{-1} M \right] \vec{a} \\ &= \text{Tr} \left(C_y^{-1} \vec{y} \vec{y}^T - C_a^{-1} \vec{a} \vec{a}^T \right) \end{aligned}$$

- expectation value χ^2_{min} (using $C_x = \left\langle \vec{x} \vec{x}^T \right\rangle - \langle \vec{x} \rangle \langle \vec{x} \rangle^T$)

$$\begin{aligned} \left\langle \chi^2_{\text{min}} \right\rangle &= \text{Tr} \left(C_y^{-1} (C_y + \langle \vec{y} \rangle \langle \vec{y} \rangle^T) - C_a^{-1} (C_a + \langle \vec{a} \rangle \langle \vec{a} \rangle^T \right) \\ &= n_y - n_a + \text{Tr} \left(C_y^{-1} \langle \vec{y} \rangle \langle \vec{y} \rangle^T - C_a^{-1} \langle \vec{a} \rangle \langle \vec{a} \rangle^T \right) \\ &= n_y - n_a \end{aligned}$$

The last step used $C_a^{-1} = M^T C_y^{-1} M$ and $M \langle \vec{a} \rangle = \langle \vec{y} \rangle$.



Summary: Least Squares Method

- formulation via the cost function ...
  - ➜ derived for linear models and explains the name “least squares”
  - ➜ easily generalizes to multi-dimensional and non-linear problems
- least squares are a distribution-free way for parameter estimates
  - ➜ requires only data and covariance matrix of the data
  - ➜ weight matrix C^{-1} must be fixed
  - ➜ approximately gaussian errors due to the central limit theorem
- for linear models
  - ➜ unbiased estimates of the true parameters
  - ➜ parameter estimates are linear combinations of the measurements
- when using the inverse of the covariance matrix as weight matrix
  - ➜ linear estimates with minimal variance
  - ➜ independent of the shape of the PDF of the fluctuations
  - ➜ goodness-of-fit criterion $\langle \chi^2_{\rm min} \rangle = N_{\rm data} - N_{\rm par} \equiv N_{\rm ndf}$

Numerical example

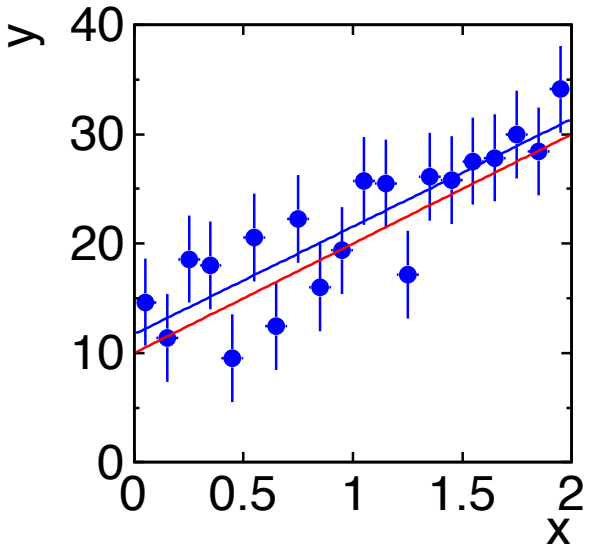
➜ *straight line fit:* $y = a_0 + a_1 x$

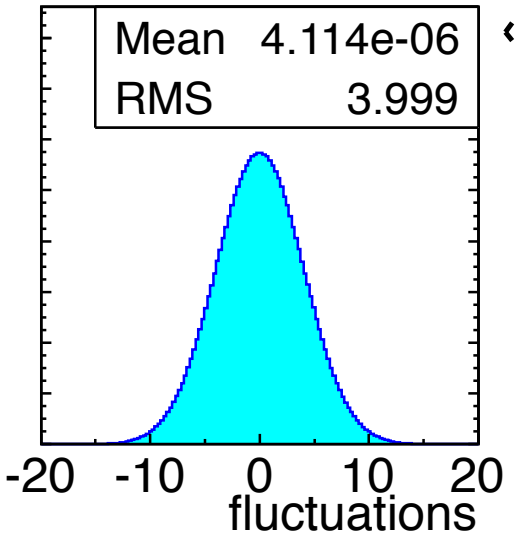
- expectation values of measurements $y(x)$: $\langle y \rangle = 10 + 10\,x$
- take 20 equidistant points in the range $0 < x < 2$
- measurements fluctuate with rms$= 4$ around the expectation value
  - ➜ gaussian distribution
  - ➜ exponential distribution
  - ➜ uniform distribution
- same covariance matrix and $\langle \chi^2_{\min} \rangle = 18$ in all cases

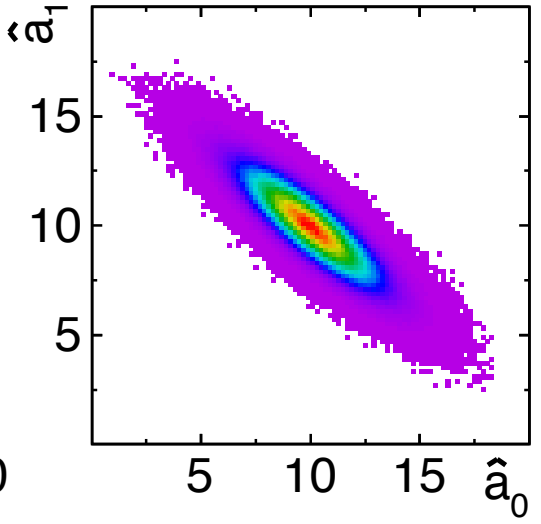
$$C(a) \approx \begin{pmatrix} 3.206 & -2.406 \\ -2.406 & 2.406 \end{pmatrix} \qquad \begin{matrix} \sigma(a_0) \approx 1.7905 \\ \sigma(a_1) \approx 1.5511 \end{matrix} \qquad \rho \approx -0.8663$$

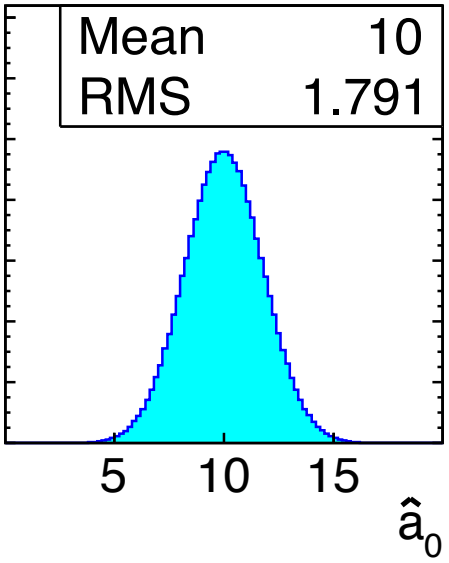
- study also poisson distributed measurements...
  - ➜ fit with correct standard deviations: $\sqrt{\langle y \rangle}$
  - ➜ fit with estimated standard deviations: $\sqrt{y}$

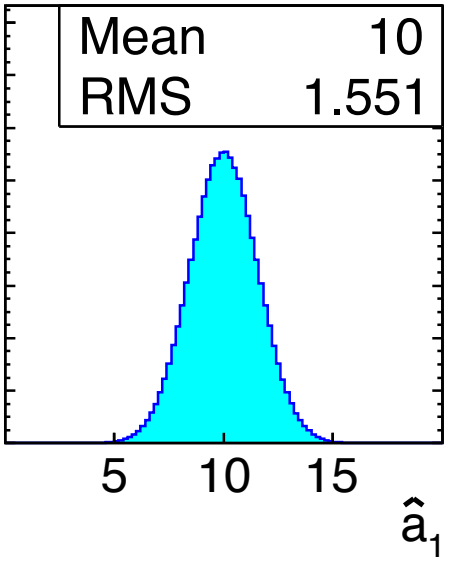
Illustration

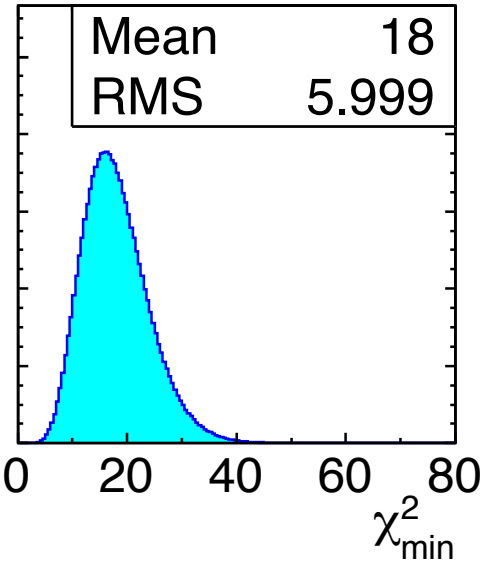












➜ *introductory remarks*

- common wisdom: least squares fits need. . .
  - ➜ gaussian fluctuations
  - ➜ sufficiently large event counts for poisson distributed data
- in the derivation of the method none of the above entered
  - ➜ only proper variance estimates were assumed
  - ➜ the variances are treated as constants in the χ^2 minimization
  - ➜ the variance estimates should not be correlated to the data

case study, keeping an eye on those points when doing fits ➜

Numerical example

➜ *determination of the lifetime of an unstable particle*

- lifetime distribution

$$\frac{dn}{dt} = \frac{1}{\mu} e^{-t/\mu} \qquad \text{with} \qquad \mu = 1\,\text{ns}$$

- MC study of test experiments with fixed number N of decays
  - ➜ histogram representation of the measurement
  - ➜ 100 bins for $0 < t < 10\,\text{ns}$
- optimal parameter estimate:

$$\hat{\mu} = \frac{1}{N} \sum_{i=1}^{N} t_i \qquad \text{for } \mu = 1\text{:} \qquad \hat{\mu} = 1 \pm \frac{1}{\sqrt{N}}$$

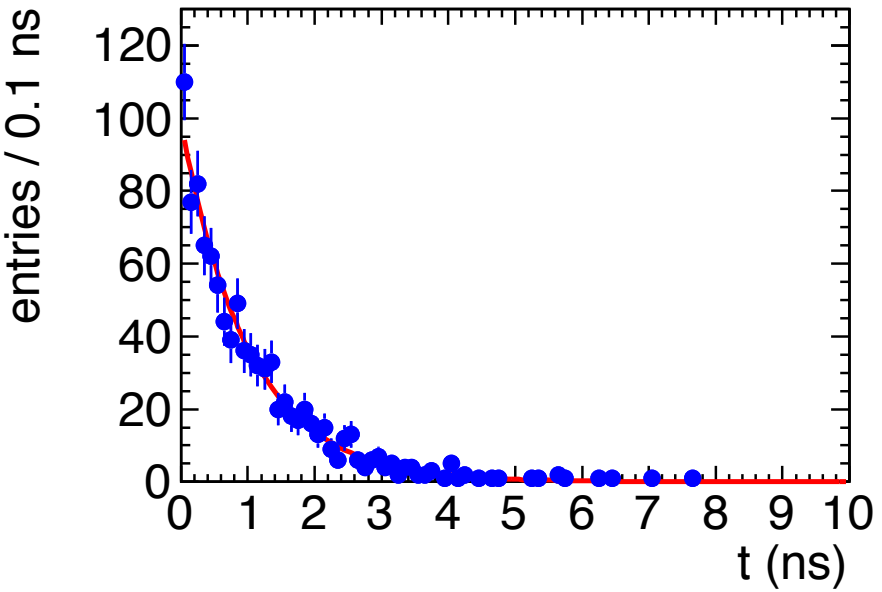
- parametric model for bin contents n_i in Least Squares fit

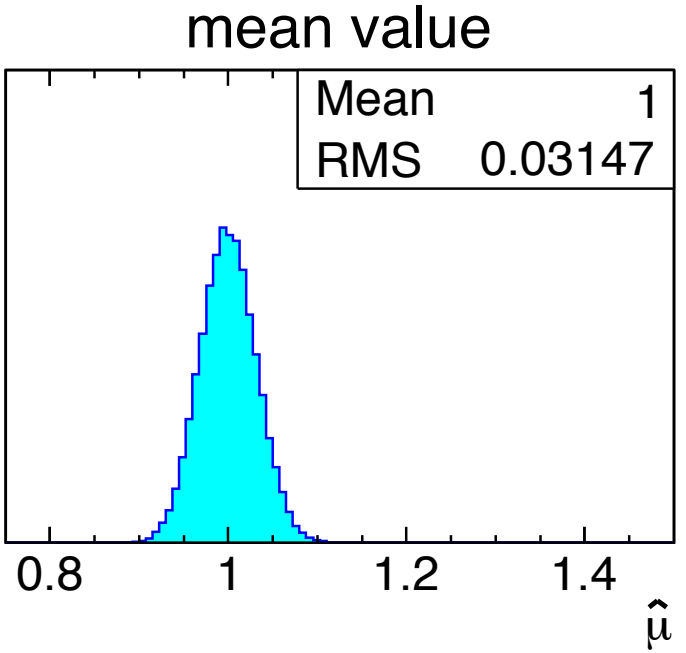
$$f_i(\mu) = N \int_{\text{bin } i} dt \; \frac{dn}{dt}$$

➜ *test different weight-assignments*

- $w_i = 1$ for all bins
  - ➜ unsophisticated but hopefully robust unweighted fit
- $w_i = 1$ for all bins with non-zero entries
  - ➜ pretend that empty bins don't have informations
- $w_i = 1/n_i$ for all bins with non-zero entries
  - ➜ use empirical variance estimates
- $w_i = 1/f_i$ for all bins
  - ➜ naive way to use the theoretical variances
- iterative fit with $w(0) = 1$ and $w_i(m) = 1/f_i(\hat{\mu}_{m-1})$ for all bins
  - ➜ proper way to use the theoretical variances
  - ➜ implements that variances must be fixed in minimization
  - ➜ weak correlation between variance estimates and data
- for comparison: simple arithmetic mean of all entries

➜ *best fit performance for $N = 1000$ events*

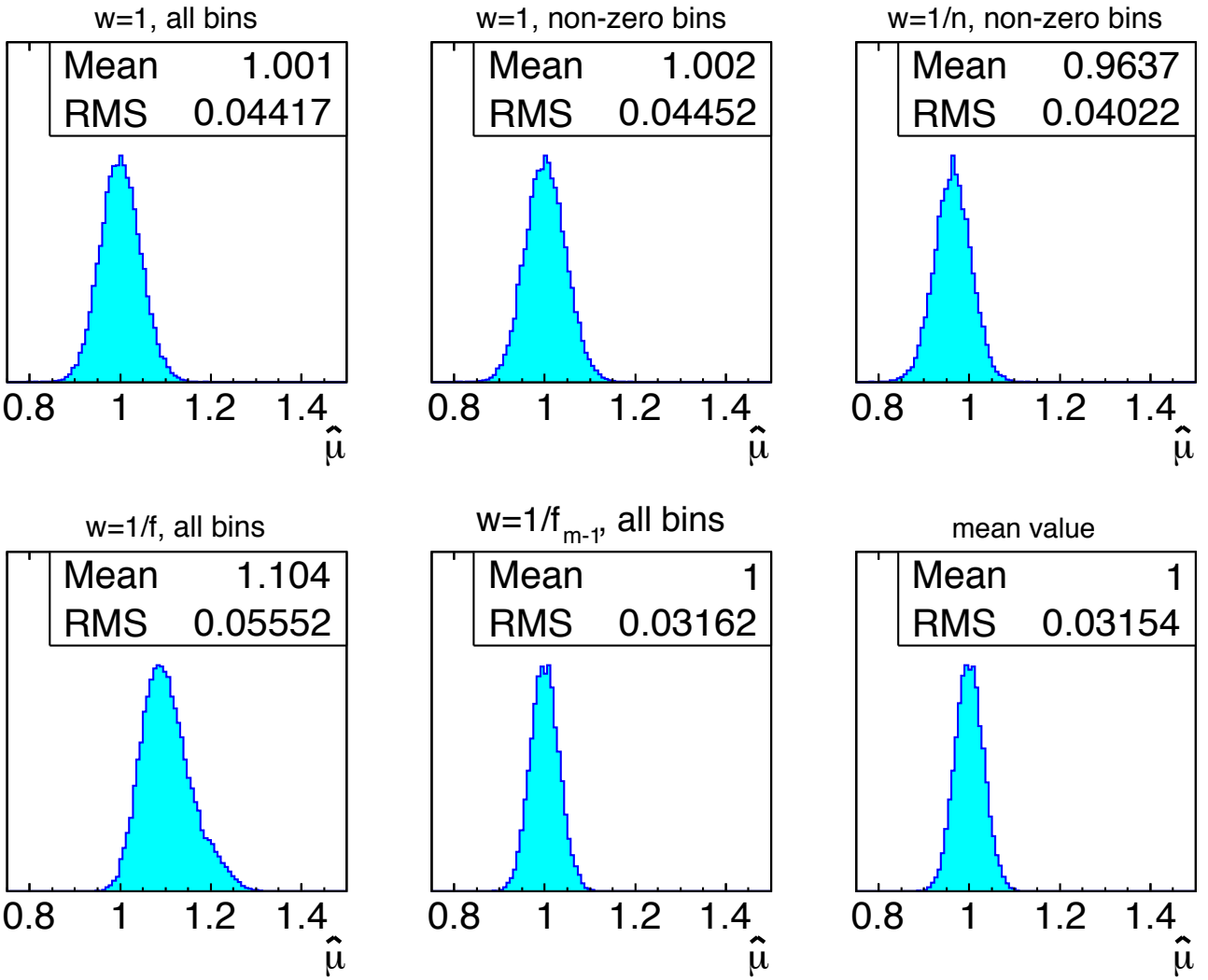




- check standard deviation and bias of fitted $\hat{\mu}$
  - ➜ as a function of available statistics
  - ➜ for the different choices of the weight function

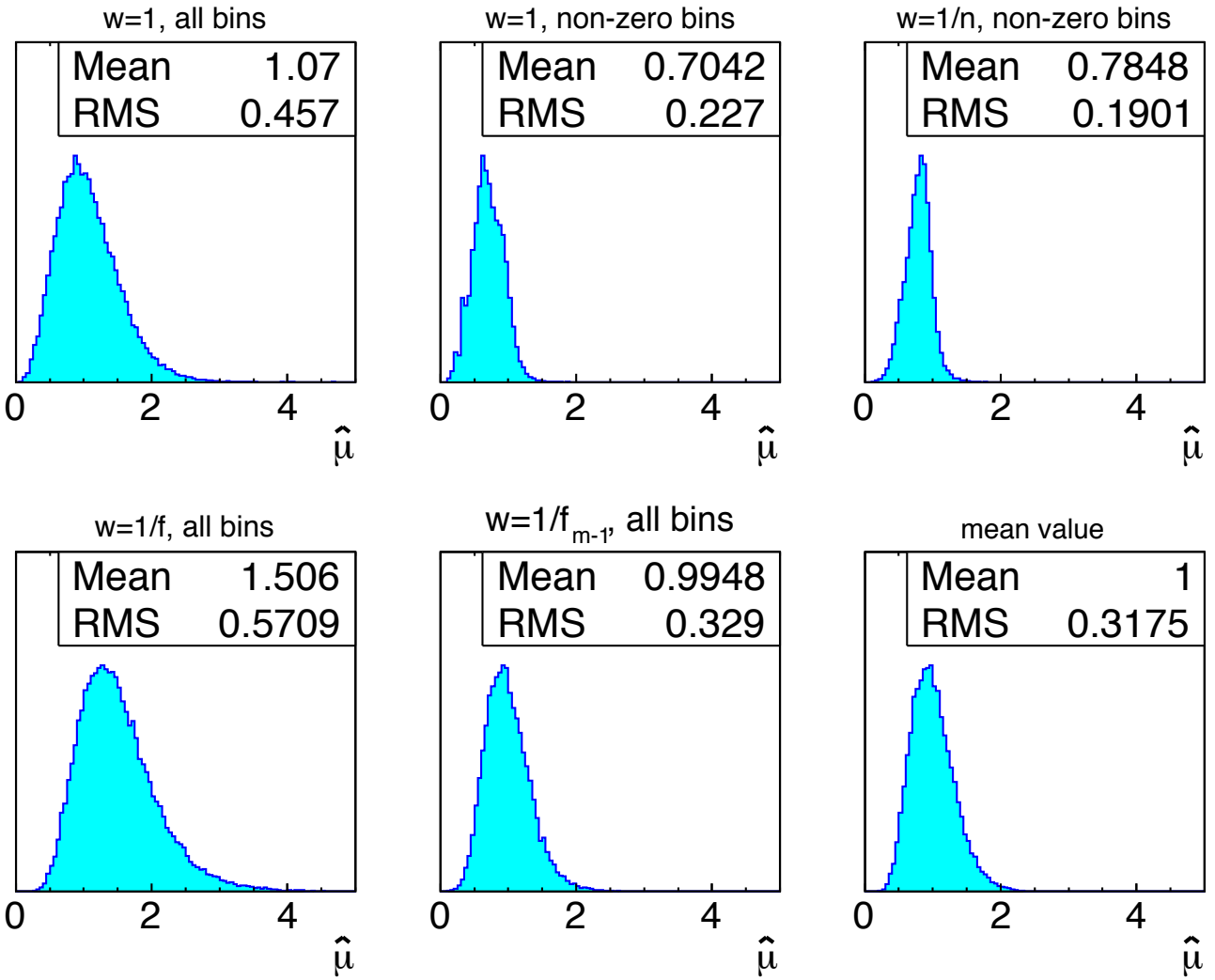
Performance comparison

➜ *parameter estimates for $N = 1000$ events*



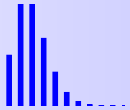
Performance comparison

➜ *parameter estimates for $N = 10$ events*



➜ *properties of different weight-assignments*

- $w_i = 1$ for all bins
  - ➜ OK, generally unbiased, but not with optimal precision
  - ➜ do not use Hessian of χ^2 function for error estimates
- $w_i = 1$ for non-zero bins
  - ➜ needless loss of information and bias at low statistics
- $w_i = 1/n_i$ for all bins with non-zero entries
  - ➜ biased – violates the least squares ansatz
- $w_i = 1/f_i$ for all bins
  - ➜ biased – violates the least squares ansatz
- iterative fit with $w(0) = 1$ and $w_i(m) = 1/f_i(\hat{\mu}_{m-1})$ for all bins
  - ➜ close to optimum (maximum likelihood fit)
  - ➜ works also for low statistics



Exercises

Consider two parameters with true values $a_0^{\text{true}} = 22$ and $a_1^{\text{true}} = 88$. Actual measurements a_0 and a_1 scatter around the true values with standard deviations $\sigma_0 = 4$ and $\sigma_1 = 8$. The correlation coefficient between them is $\rho = C_{01}/\sqrt{C_{00}\,C_{11}} = -0.5$. From a_0 and a_1 the value $b = a_1/a_0$ shall be determined.

1) Determine analytically for a pair (a_0, a_1) the bias corrected value of b and it's uncertainty.
2) Assuming gaussian fluctuations, generate pairs (a_0, a_1) and histogram the ratio a_1/a_0, the bias corrected ratio and the estimate of the uncertainty.

- If x_1 and x_2 are independent random numbers with mean value zero and unit variance, a pair of correlated random numbers (y_1, y_2) is obtained by $(y_1 = x_1, y_2 = x_1\rho + x_2\sqrt{1-\rho^2})$.

➜ *uncertainty of* $b = a_1/a_0$

$$\sigma^2(b) = \left(\frac{\partial b}{\partial a_0}\right)^2 C_{00} + 2\frac{\partial b}{\partial a_0}\frac{\partial b}{\partial a_1} C_{01} + \left(\frac{\partial b}{\partial a_1}\right)^2 C_{11}$$

result:

$$\sigma^2(b) = \frac{a_1^2}{a_0^4}\,16 + \frac{a_1}{a_0^3}\,32 + \frac{1}{a_0^2}\,64$$

➜ *bias (to be subtracted)*

$$B = \frac{1}{2}\left(\frac{\partial^2 b}{\partial a_0^2} C_{00} + 2\frac{\partial^2 b}{\partial a_0 \partial a_1} C_{01} + \frac{\partial^2 b}{\partial a_1^2} C_{11}\right)$$

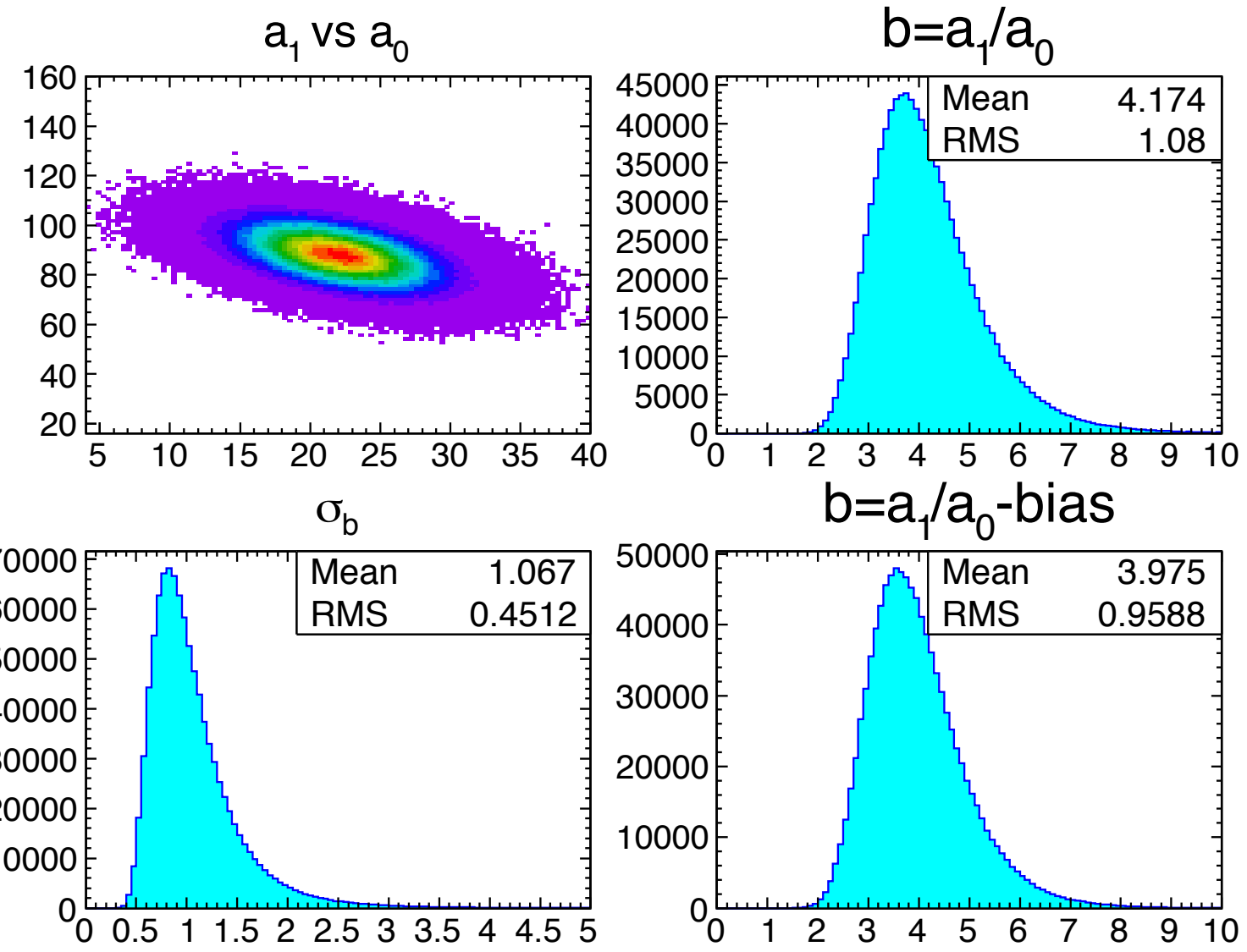
result:

$$B = \frac{a_1}{a_0^3}\,16 + \frac{1}{a_0^2}\,16$$

➜ *simulation of a single measurement*

```
double x1 = rndm.Gaus();
double x2 = rndm.Gaus();
double a0 = av0 + sd0*x1;
double a1 = av1 + sd1*(rho*x1+x2*sqrt(1.-rho*rho));
double b  = a1/a0;
double db = sqrt(C00*a1*a1/(a0*a0)-2*C01*a1/a0+C11)/a0;
double B  = (C00*a1/a0 - C01)/(a0*a0);
```

Numerical results



Consider y_k, $k = 1, 2, \ldots, 10$ poisson-distributed measurements with expectation values $\mu_k = a \cdot k$. The parameter a shall be determined by a least squares fit. Consider the following χ^2 functions:

$$\chi^2 = \sum_{k=1}^{10} w_k \left(y_k - ak\right)^2 \quad \text{with} \quad w_k \in \left\{1, \frac{1}{y_k}, \frac{1}{a\,k}, \frac{1}{a_{n-1}k}, \frac{1}{k}\right\}$$

1) Determine analytically the best fit values $\hat{a}$ for the various weight functions
2) Generate toys for $a \in \{1, 2, 4\}$ and determine the distribution of $\hat{a}$ for the various options.

➜ *analytical calculation*

$$\chi^2 = \sum_{k=1}^{10} w_k (y_k - ak)^2 = \sum_{k=1}^{10} (w_k y_k^2 - 2aw_k y_k k + a^2 w_k k^2)$$
$$\equiv S_{wyy} - 2aS_{wyk} + a^2 S_{wkk}$$

one obtains:

$$\chi^2(w_k = 1) = S_{yy} - 2aS_{yk} + a^2 S_{kk}$$
$$\chi^2(w_k = 1/y_k) = S_y - 2aS_k + a^2 S_{kk/y}$$
$$\chi^2(w_k = 1/ak) = (1/a)S_{yy/k} - 2S_y + aS_k$$
$$\chi^2(w_k = 1/k) = S_{yy/k} - 2aS_y + a^2 S_k$$

The iterated weight function $w_k = 1/(a_{n-1}k)$ gives the same estimate as $w_k = 1/k$, since the χ^2 minimum is not affected by a global scale factor.

➜ *best fit $\hat{a}$ values*

$$\hat{a}(w_k = 1) = \frac{S_{yk}}{S_{kk}}$$

$$\hat{a}(w_k = 1/y_k) = \frac{S_k}{S_{kk/y}}$$

$$\hat{a}(w_k = 1/ak) = \sqrt{\frac{S_{yy/k}}{S_k}}$$

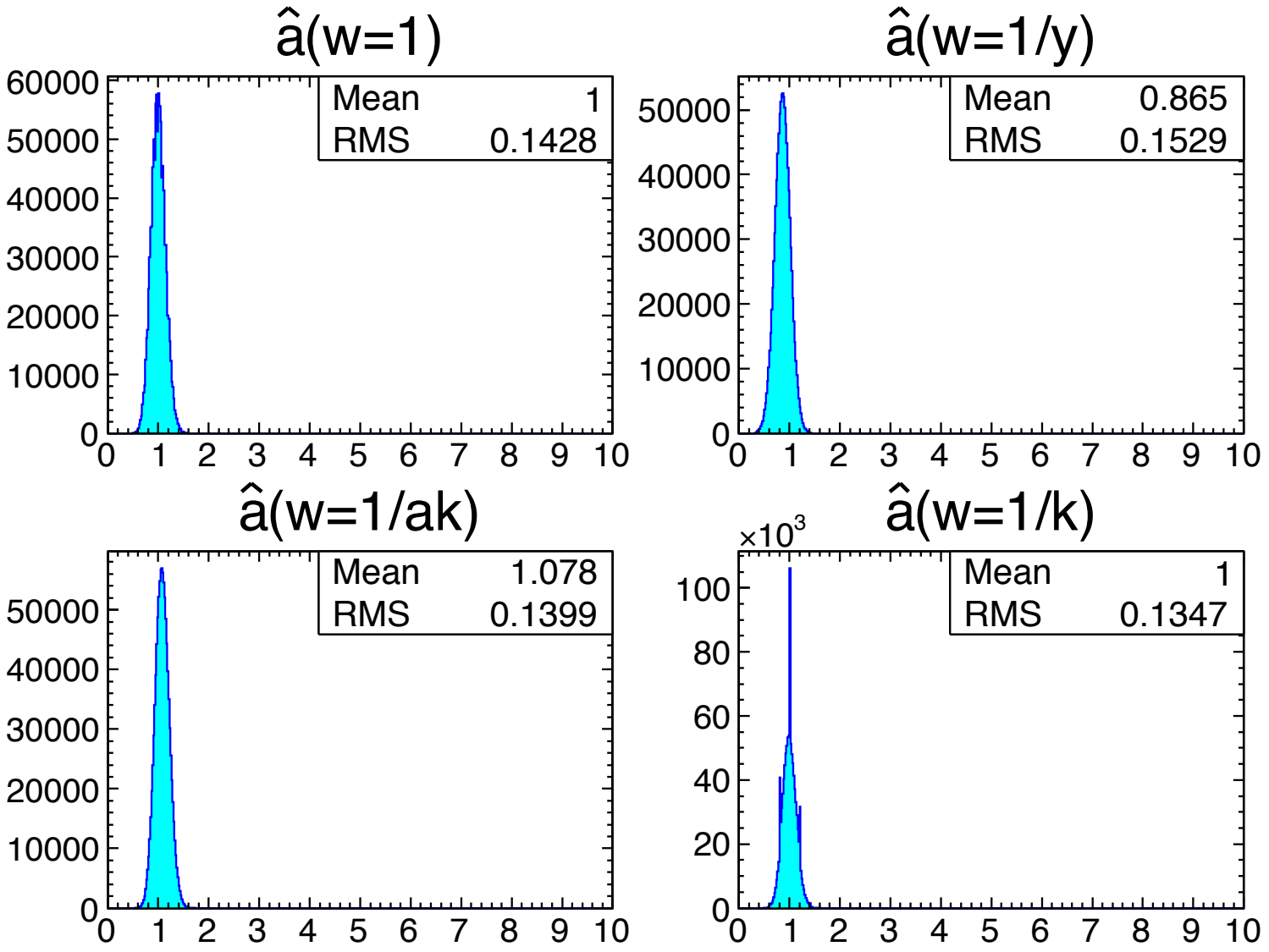
$$\hat{a}(w_k = 1/k) = \frac{S_y}{S_k}$$

All estimators have the correct dimension $[\hat{a}] = [y]/[k]$, which is required for $y = ak$. The first and the last estimator are expected to be unbiased, the two middle ones might have problems.

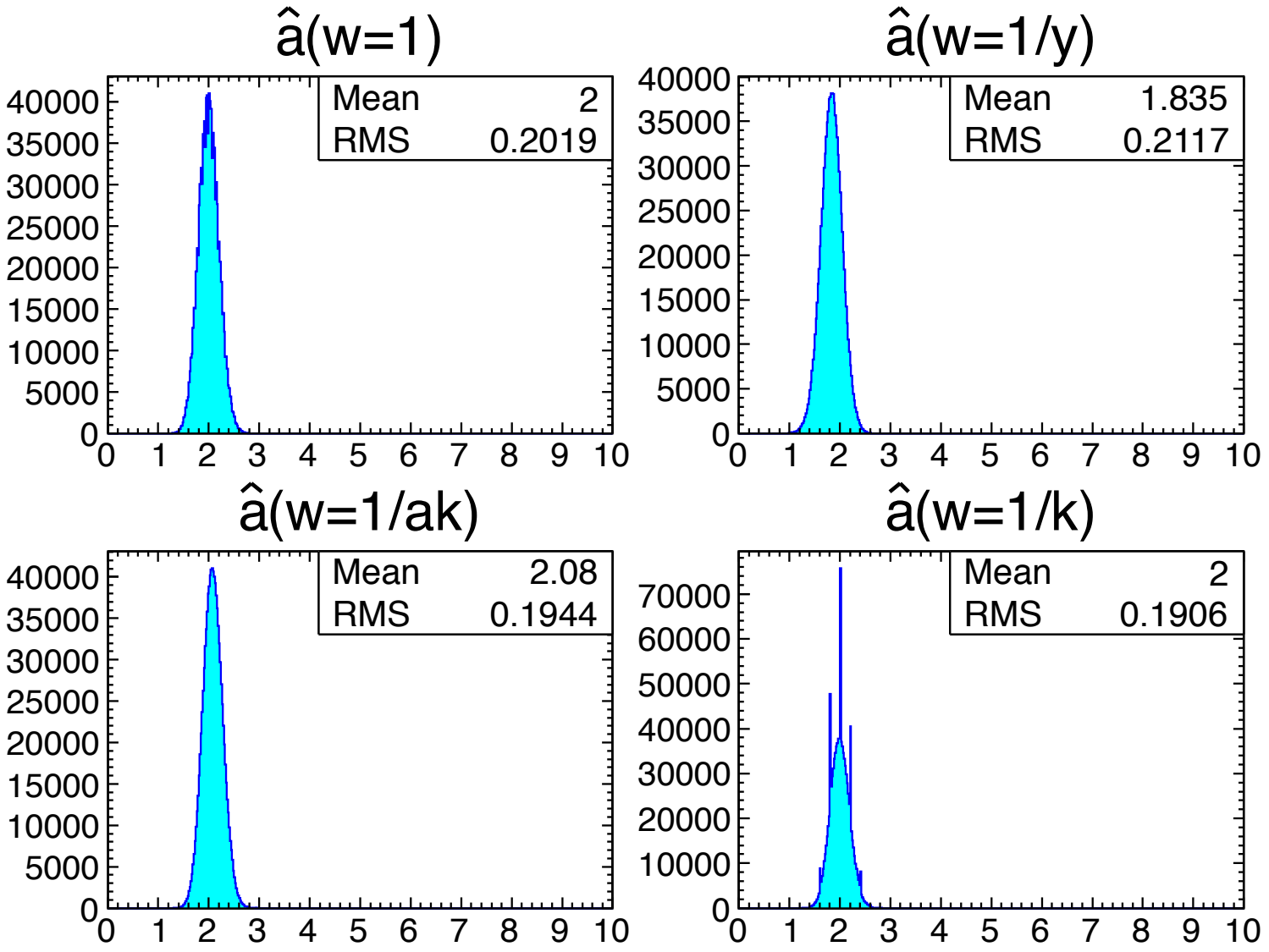
➜ *simulation of a single measurement*

```
Sy =Sk = Syk = Skk = Syyk = Sk1 = Skky = 0.;
for(int n=1; n<=10; ++n) {
  double k = double(n);
  double y = rndm.Poisson(a*k);
  Sy += y;
  Sk += k;
  Syk += y*k;
  Skk += k*k;
  Syyk += y*y/k;
  if(y>0.) Sk1  += k;
  if(y>0.) Skky += k*k/y;
}
ha0->Fill(Syk/Skk);
ha1->Fill(Sk1/Skky);
ha2->Fill(sqrt(Syyk/Sk));
ha3->Fill(Sy/Sk);
```

Numerical results for $a = 1$



Numerical results for $a = 2$



Numerical results for $a = 4$