# Statistics for Data Analysis

PSI Practical Course 2014

Niklaus Berger

Physics Institute, University of Heidelberg

Emmy Noether-Programm
Deutsche Forschungsgemeinschaft
DFG

# Overview

You are going to perform a data analysis:
Compare measured distributions to theoretical predictions

## Tools for data analysis:

- Probability density functions,
- Histograms,
- Fits,
- Errors

This is not a statistics course; no proofs, not too many details
(Attend C. Grab's or my/Oleg Brandt's course for more...)
Thanks to C. Grab for most of the material

# Probability vs. Statistics

Probability: From theory to data

Start with a well-defined problem,
calculate all possible experimental outcomes

Statistics: From data to theory

Inverse problem: Start with (messy) data,
deduce rules, laws: Data Analysis

Parameter estimation: Determine parameter & error
in an efficient and unbiased way

Hypothesis testing: agreement, confidence...

# Probability Density Functions

# Probability and density function

Define:

Probability = #success / #trials

(classical, frequentist sense - think of throwing dice)

Experiment measures observable x many times -

results will be distributed according to some

Probability distribution:

- Individual measurements fluctuate because of uncontrolled random parameters
  e.g. noise in a voltage measurements

- The underlying physics can be probabilistic
  e.g. particle lifetimes, scattering

Probabilty distributions can be discrete or continuous (dice/lifetime)

# Probability density function (pdf)

- Repeat experiment measuring a single continuous variable x

- The probability to measure x in the interval (x, x+dx) is given by the probability density function (pdf) f(x):

$$f(x) = \lim_{dx \to 0} \frac{P(x \leq result \leq x + dx)}{dx}$$

- P is a measure of how often a value of x occurs in a given interval

$$P(x_1 \leq x \leq x_2) = \int_{x_1}^{x_2} f(x)\, dx$$

- The pdf is positive definite and normalised to 1:

$$\int_{x_{min}}^{x_{max}} f(x')\, dx' = 1$$

# Cumulative distribution function

Cumulative distribution function F(x), also known as probability distribution function

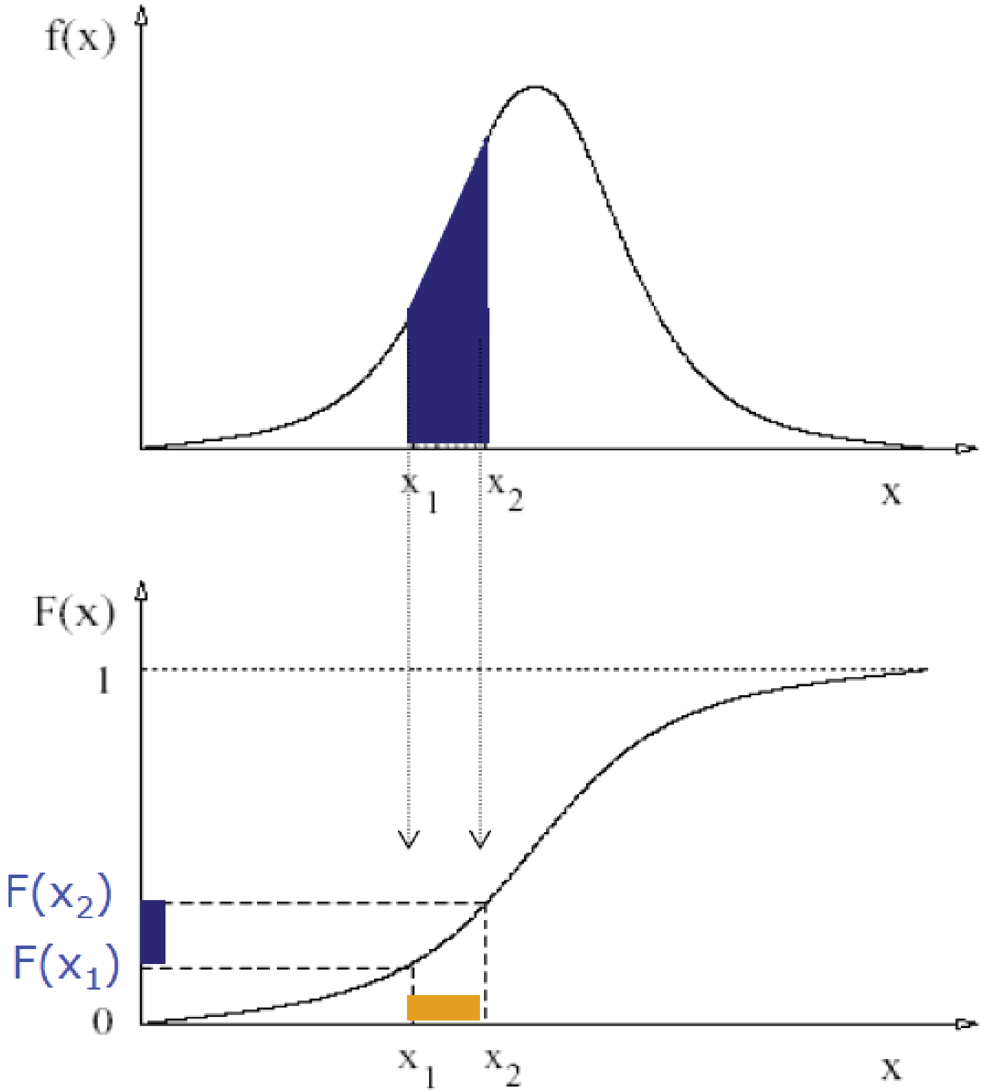
- F(x) is the probability that in am measurement, we find a value less than x
- F(x) is a continuously non-decreasing function
- $F(-\infty) = 0$, $F(\infty) = 1$
- F(x) is dimensionless
- related to the pdf f(x) by:

$$F(x) = \int_{x_{min}}^{x} f(x')\,dx'$$

- and for well-behaved distributions:

$$f(x) = \frac{dF(x)}{dx}$$

# Relation: pdf f(x) and cdf F(x)




$$f(x) = \frac{dF(x)}{dx}$$

$$F(x) = \int_{x_{min}}^{x} f(x')\,dx'$$

$$P(x_1 \leqslant x \leqslant x_2) = \int_{x_1}^{x_2} f(x')\,dx' = F(x_2) - F(x_1)$$

# Properties of distributions

- Expectation value = mean value

$$E[x] = \int_{x_{min}}^{x_{max}} x\, f(x)\, dx = \langle x \rangle = \mu$$

- Variance $\sigma^2$ = square of the standard deviation = measure of the variations of x around the mean value E[x]

$$V[x] = E[(x-\mu)^2] = \int_{x_{min}}^{x_{max}} (x-\mu)^2 f(x)\, dx = \sigma^2 = \langle (x-\mu)^2 \rangle = \langle x^2 \rangle - \mu^2$$

- Note: $\sigma$ measures how spread-out the distribution is, not how accurate the mean is determined

# Properties of distributions

- True mean and variance: both unknown...

$$E[x] = \int_{x_{min}}^{x_{max}} x\, f(x)\, dx = \langle x \rangle = \mu$$

$$\sigma^2 = \int_{x_{min}}^{x_{max}} (x-\mu)^2 f(x)\, dx$$

- For discrete measurements: $\bar{x}$ is an unbiased estimator for the mean

$$\bar{x} = \frac{1}{N} \sum_i x_i$$

$$E[\bar{x}] = \mu$$

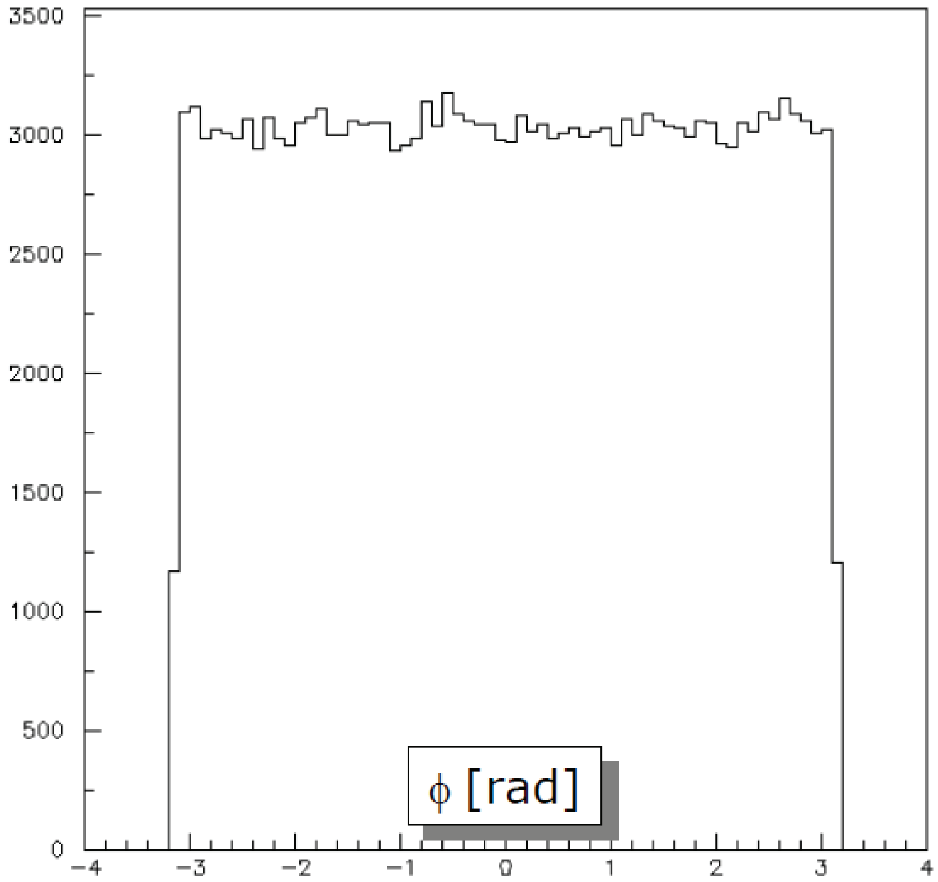
- and the sample variance $s^2$ is an unbiased estimator for $\sigma^2$

$$s^2 = \frac{1}{N-1} \sum_i (x_i - \bar{x})^2$$

$$E[s^2] = \sigma^2$$

# Examples of
# Probability Density Functions

# Uniform distribution




- Example: Polar angle distribution of muons in $e^+e^- \rightarrow \mu^+\mu^-$

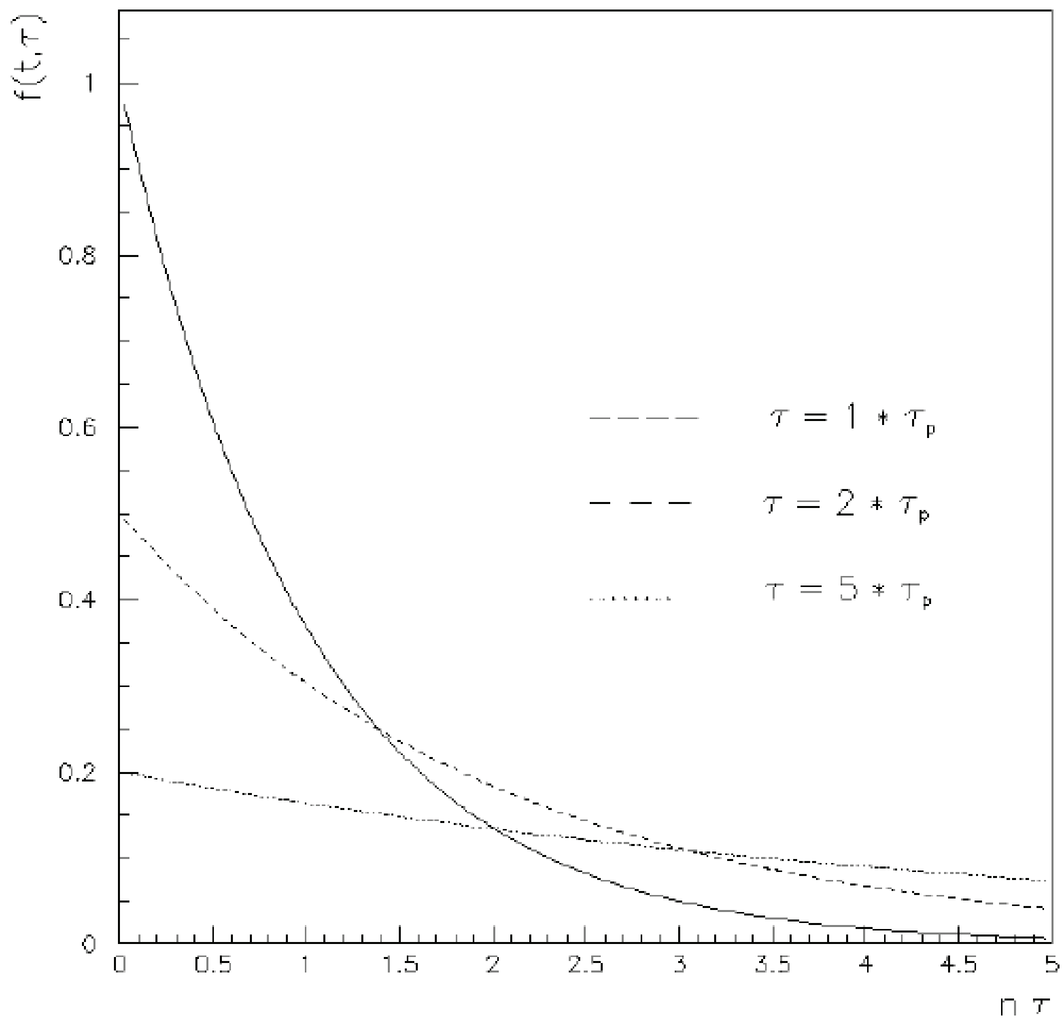
$$f(x;\ \alpha,\beta) = \begin{cases} \dfrac{1}{\beta-\alpha} & \alpha \le x \le \beta \\ 0 & \text{otherwise} \end{cases}$$

$$E[x] = \frac{1}{2}(\alpha + \beta)$$

$$V[x] = \frac{1}{12}(\beta - \alpha)^2$$

# Exponential distribution

- Example: Lifetime of the pion, muon...

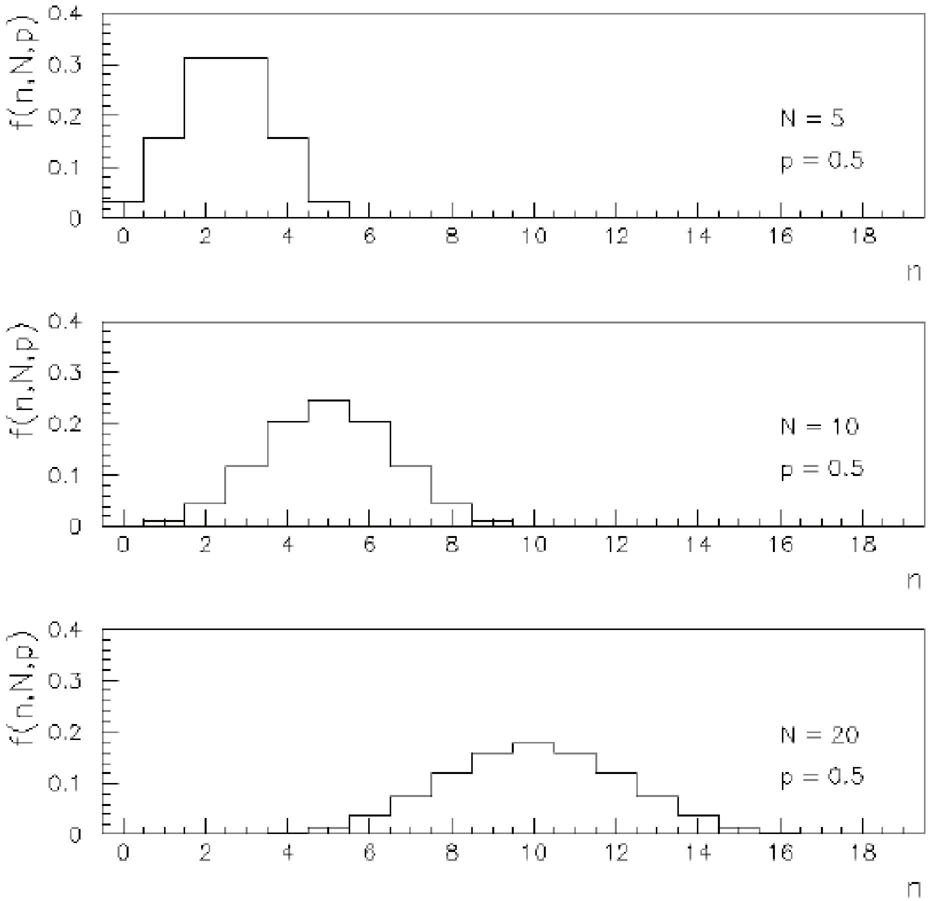



$$f(t;\ \tau) = \frac{1}{\tau} e^{-\frac{t}{\tau}}$$

$$E[t] = \tau$$
$$V[t] = \tau^2$$

# Binomial distribution

- N independent, fixed trials; probability for success = p
- Distribution of n successful outcomes in N trials
- Example: Throwing a coin/dice, chance of obtaining n heads, sixes in N throws)



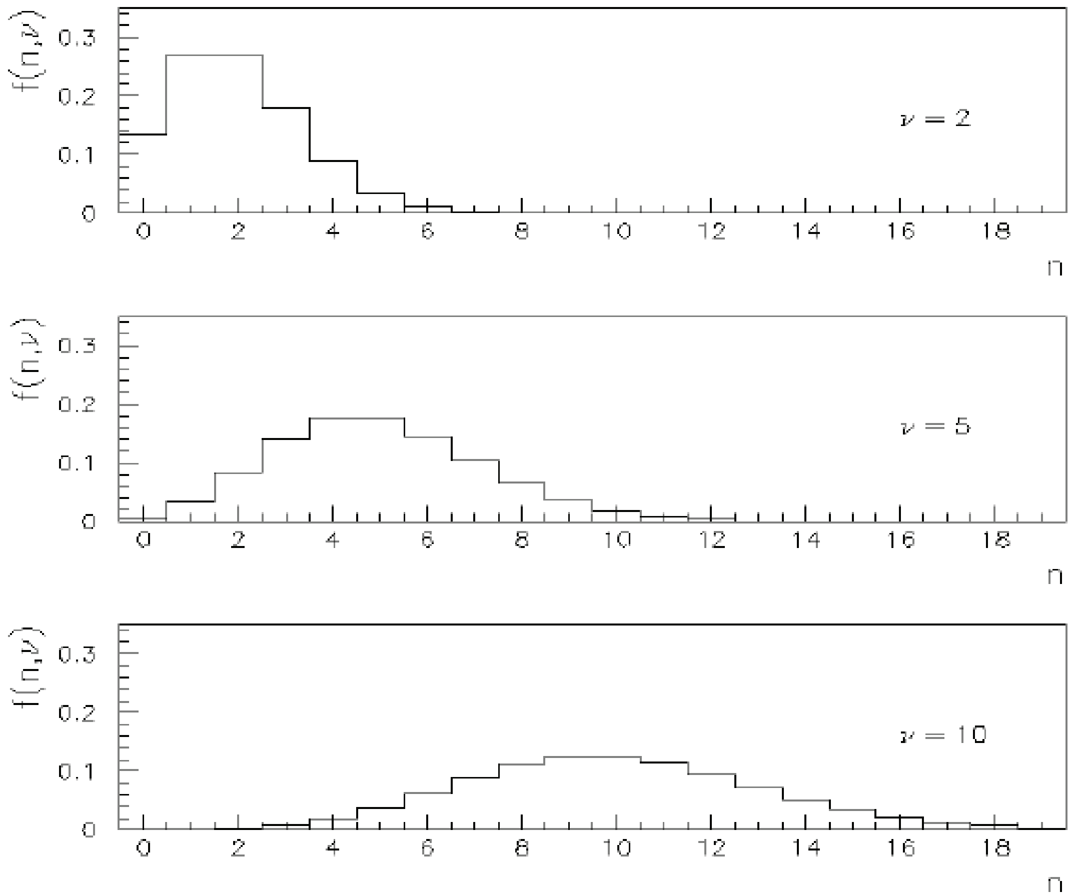

$$f(n;\ N,p) = \frac{N!}{n!(n-N)!}\,p^{n}(1-p)^{N-n}$$

$$E[n] = Np$$

$$V[n] = Np(1-p)$$

# Poisson distribution

- Limit of the binomial distribution for many trials, rare events
- $N \to \infty$, $p \to 0$ with $Np = \nu$ finite




$$f(n;\ \nu) = \frac{\nu^n}{n!} e^{-\nu}$$

$$E[n] = \nu$$

$$V[n] = \nu$$

# Poisson distribution

- Example for the Poisson distribution is:

  P(n;ν) = Probability of observing a number of n independent events in time interval t, when the average counting rate is μ; (expected number of events ν = μ t):

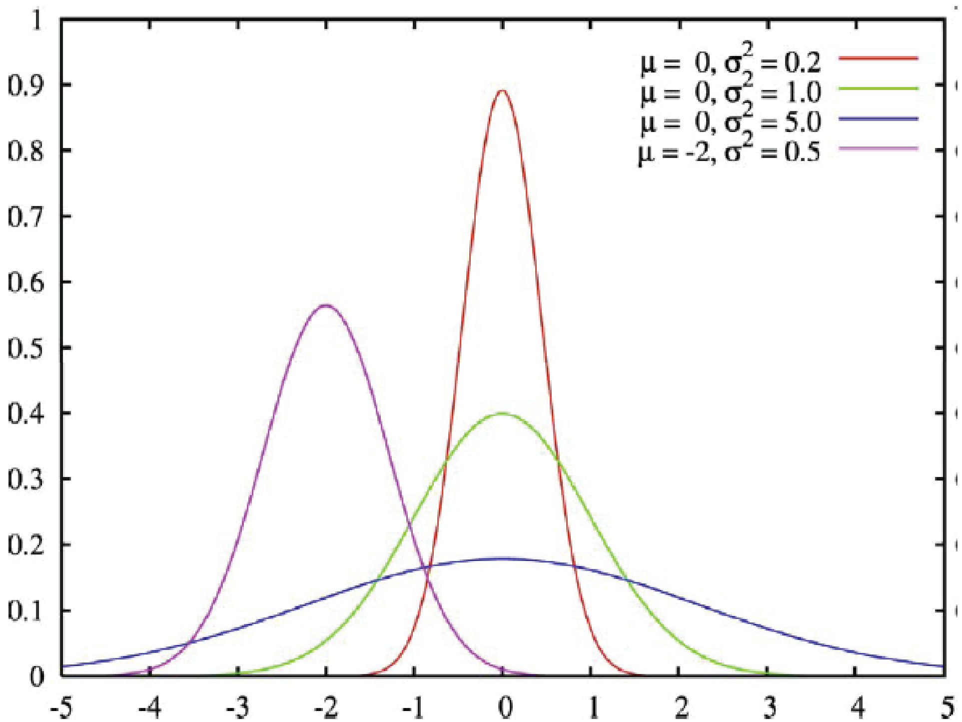
$$P(n;\nu) = \frac{(\nu)^n}{n!} e^{-\nu}$$

- Note: The variance of the Poisson distribution is equal to the expectation value ν:

  This is the origin of the formula (N ± √N) used for statistical errors when counting events during fixed intervals

# Gaussian distribution

- Also known as normal distribution
- Most important pdf...




$$f(x;\mu,\sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

$$E[x] = \mu \qquad V[x] = \sigma^2$$

- Can convert any Gaussian to standard distribution $G(\mu = 0, \sigma = 1)$ by variable transformation:

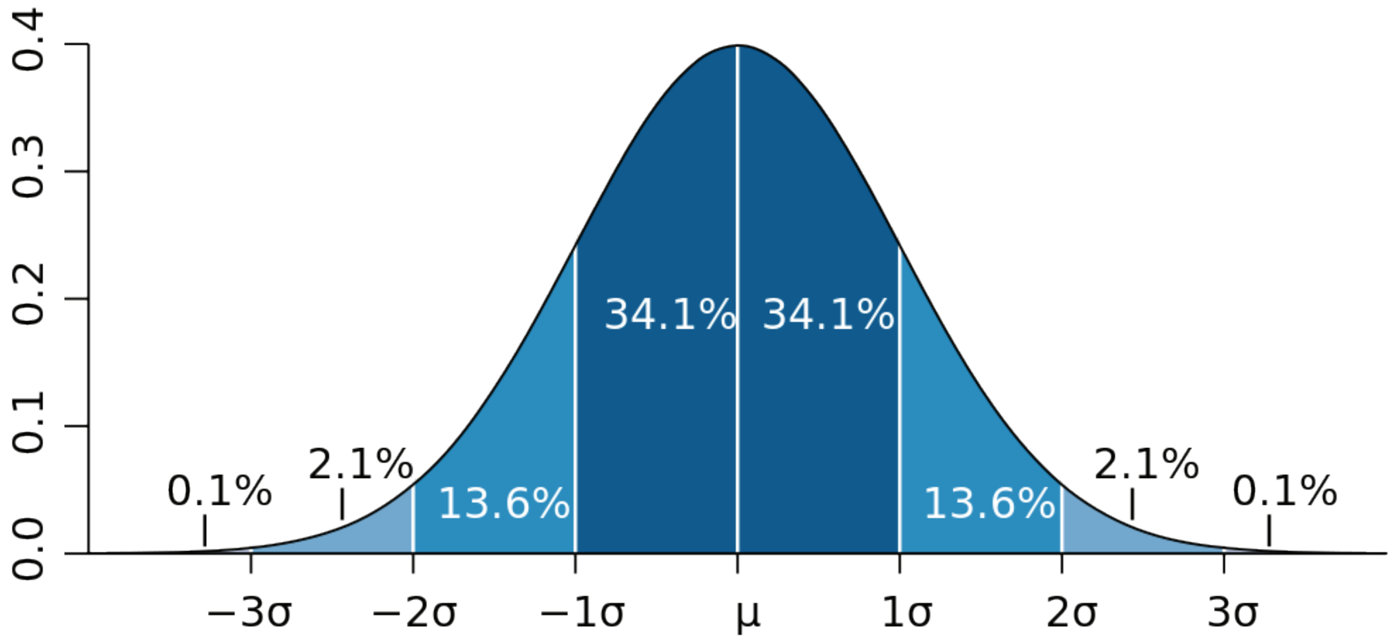
  $x' = (x - \mu)/\sigma$

# Central limit theorem

- Sum of n independent random variables $x_i$ is Gaussian distributed for $n \to \infty$
- Individual distributions do not matter!

# Properties of the Gaussian distribution

- Symmetric around $x = \mu$
- $\sigma$ characterises the width
- Height of the curve at $x = \mu \pm \sigma$ is $1/\sqrt{e}$ of the height at $x = \mu$
- $\sigma$ is roughly half the width at half the height
- Integrate area: see below;
  In 1D: $\pm 1\sigma$ : 68% (2 in 3)
  $\pm 2\sigma$ : 95%
  $\pm 3\sigma$ : 99.5%

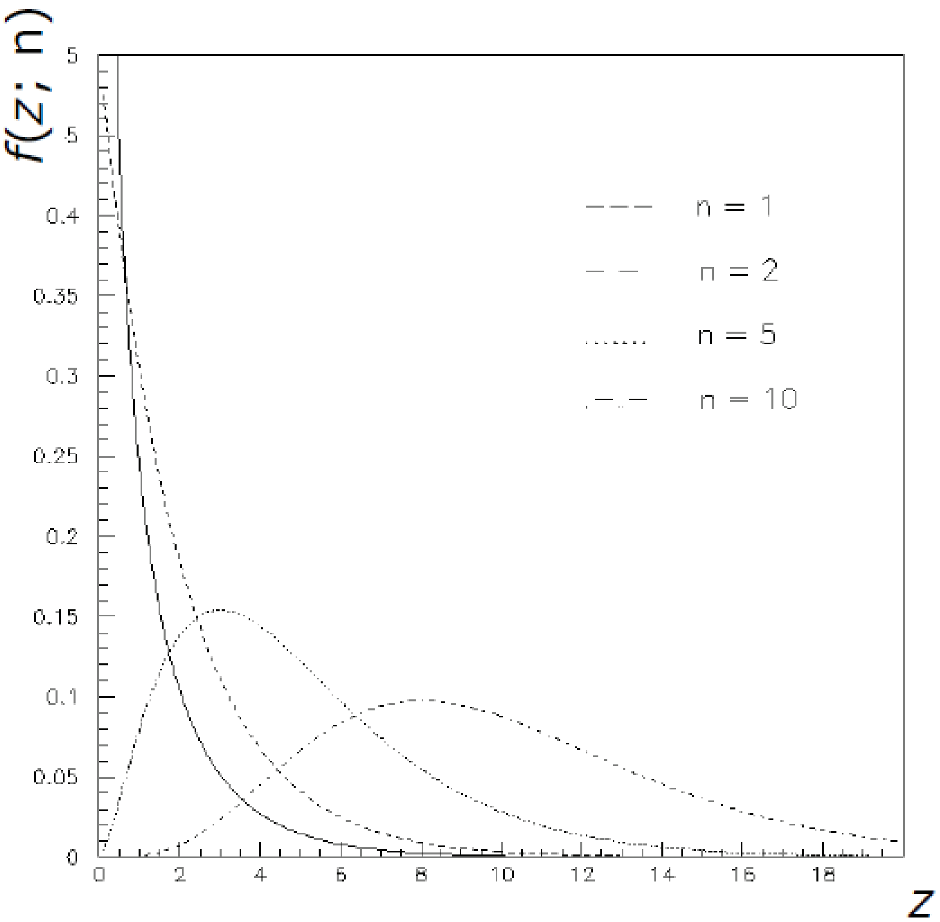


# $\chi^2$ distribution

- If $x_1 \ldots x_n$ are independent, Gaussian distributed variables with mean μ and variance σ, then

$$z = \sum_n ((x_i - \mu)/\sigma)^2$$

is distributed according to the $\chi^2$ distribution



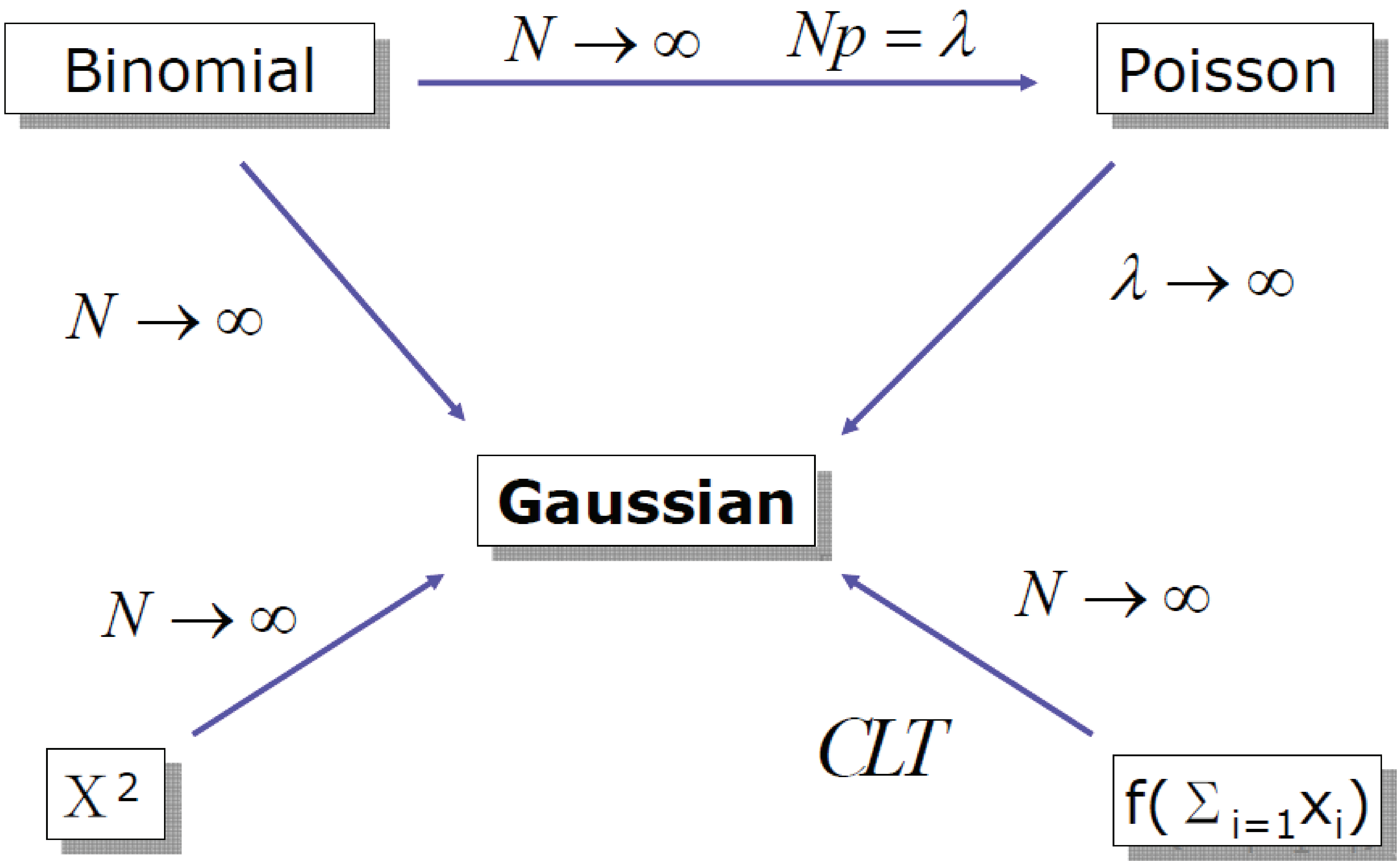

$$f(z;\ n) = \frac{1}{2^{n/2}\Gamma(n/2)} z^{n/2-1} e^{-z/2} \quad ; \quad n = 1, 2, \ldots$$

$$E[z] = n$$

$$V[z] = 2n$$

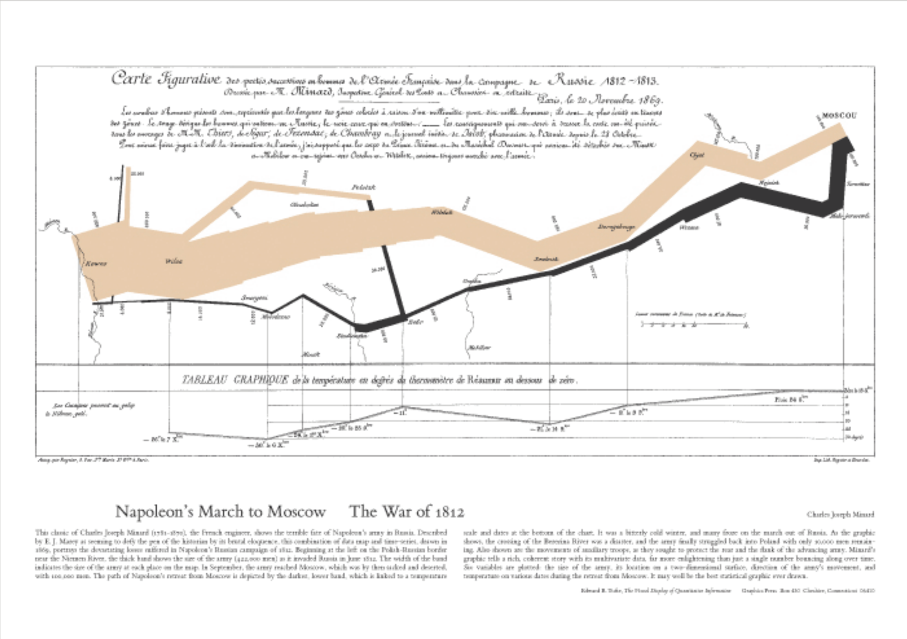
Mean is = n =
number of degrees of freedom

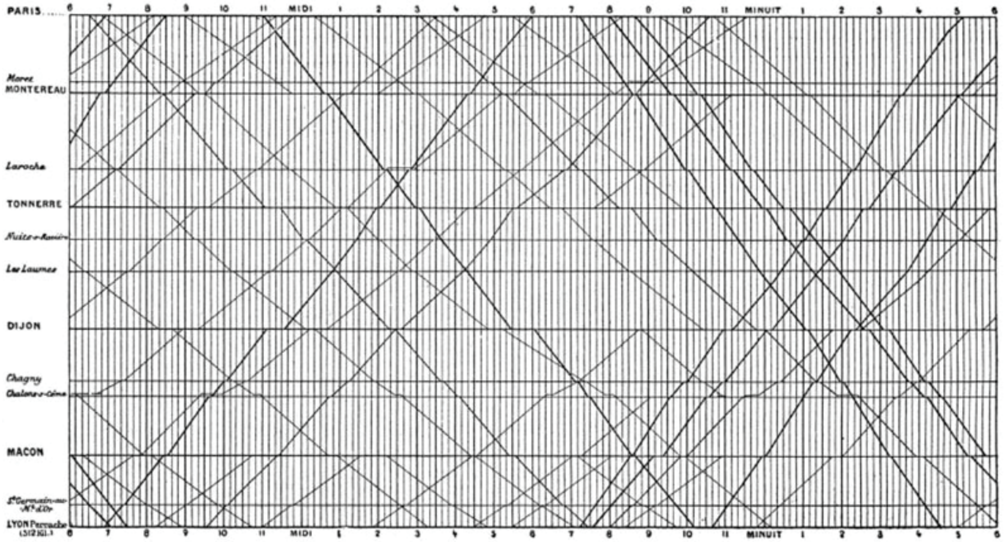
# Relations between distributions



# Histograms

# Data presentation



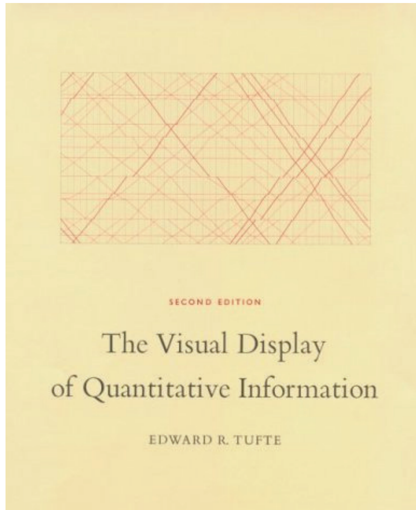




Many different ways to display quantitative data

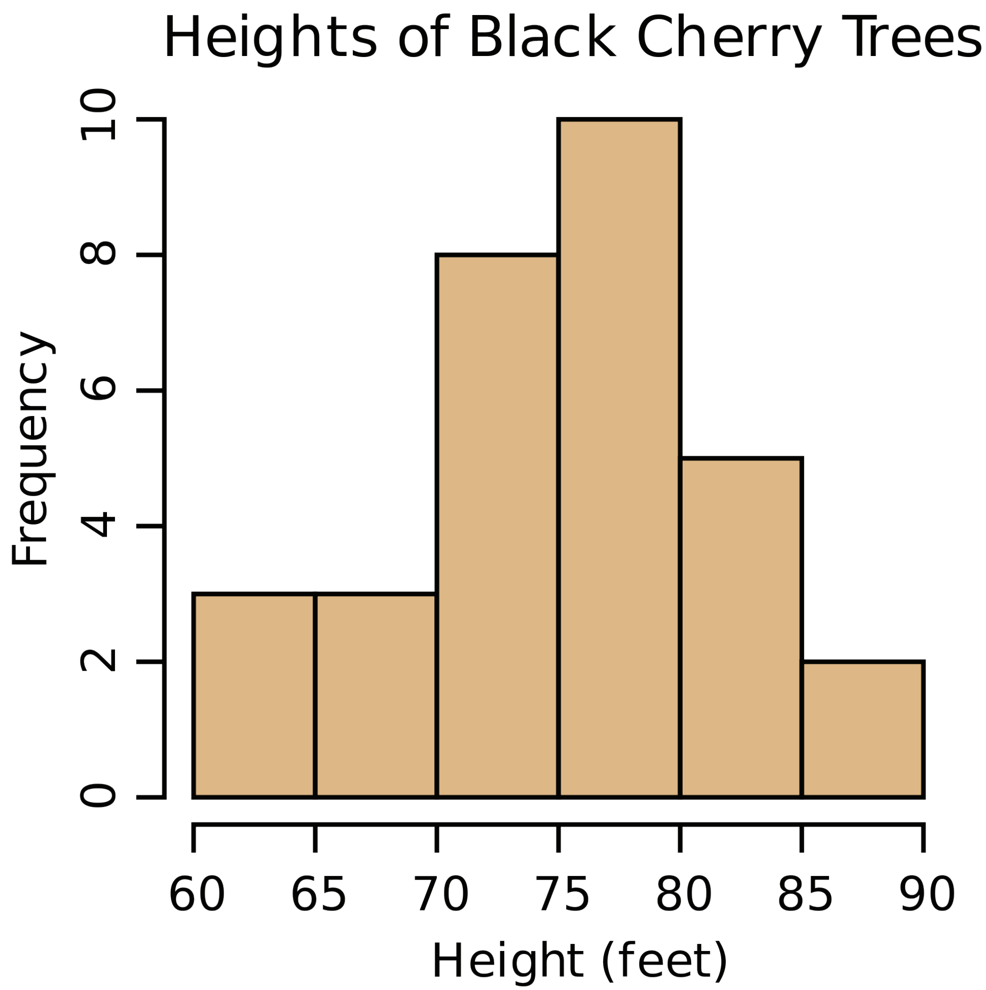
- Ideographs,
- Pie charts,
- Tables,
- Frequency polygons
- Histograms

Think about what you do…

Literature: Tufte



# Histograms

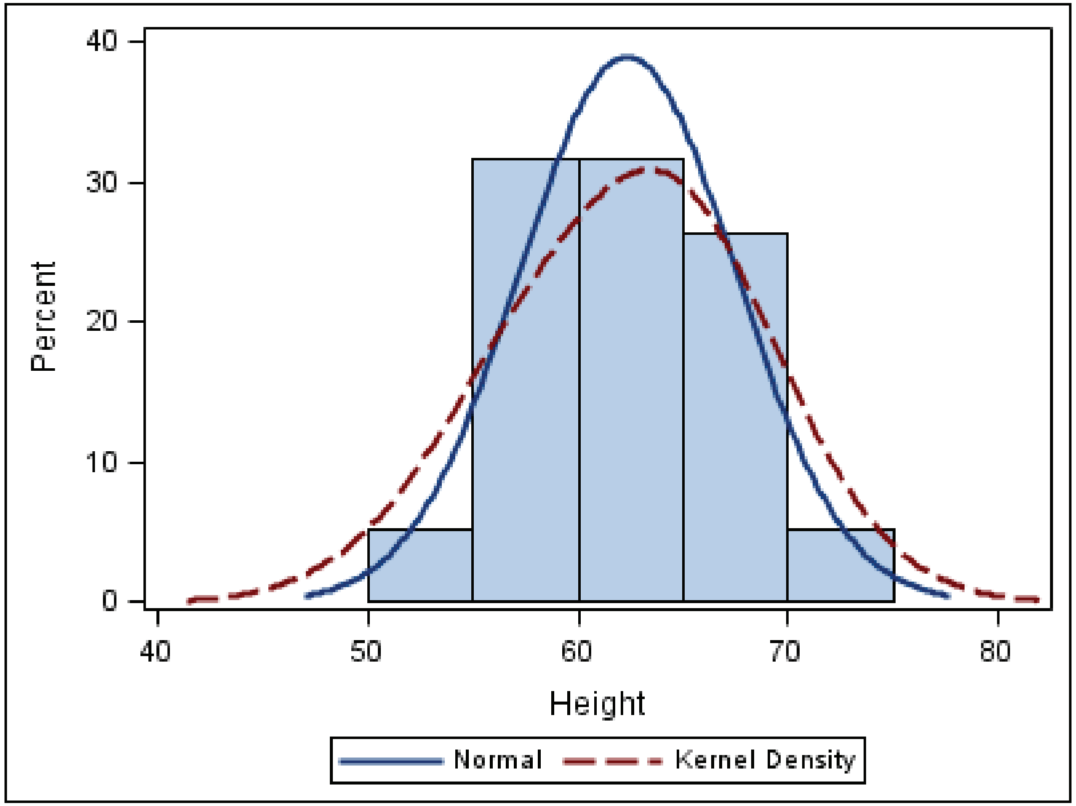



Discrete outcomes of an experiment $x_1 \ldots x_n$

- Fill into bins of a histogram
- Shape of the histogram will approximate underlying distribution:
  Can compare to (smooth) expectation/ theory curve
- Use care in choosing bin sizes, number of bins…

# Histograms

- For many entries N, histogram should approximate the probability density function
  Interpret histogram as an approximation to an underlying pdf




- What does "approximate" mean here?
- Have to look at:
  - Errors of a histogram entry
  - Normalized histograms
  - Mean values - useful or not?

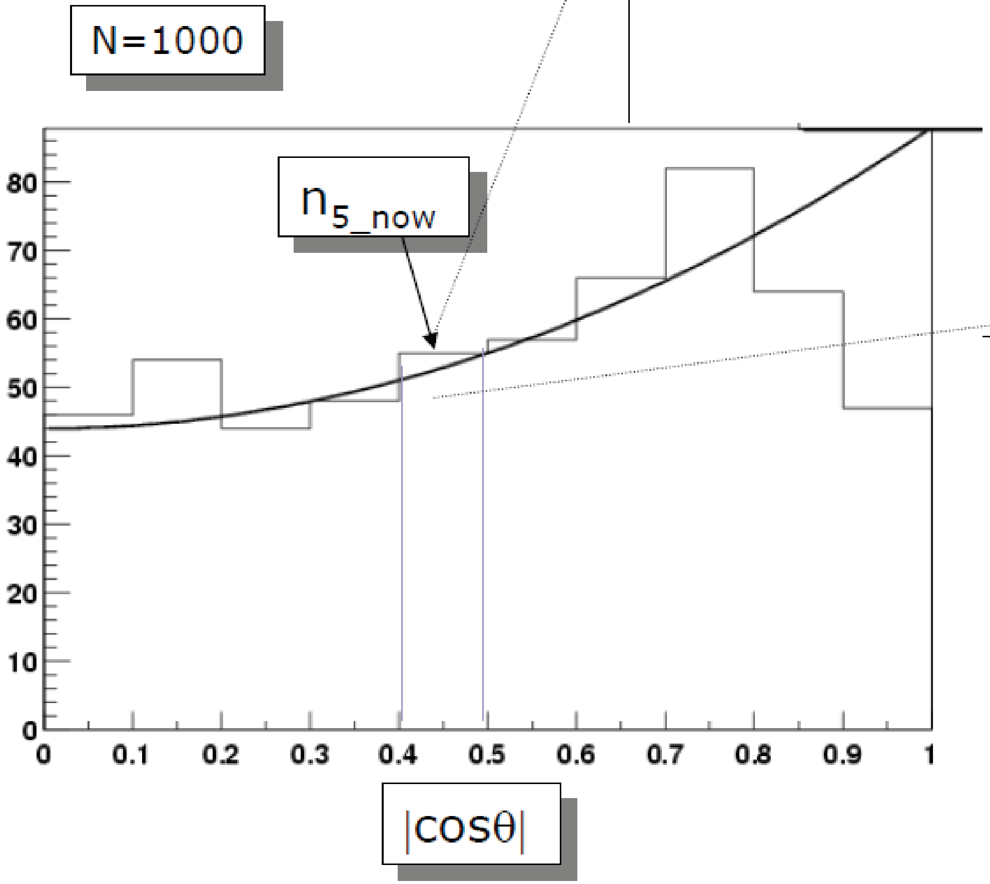
# Histogram: Interpretation and Errors

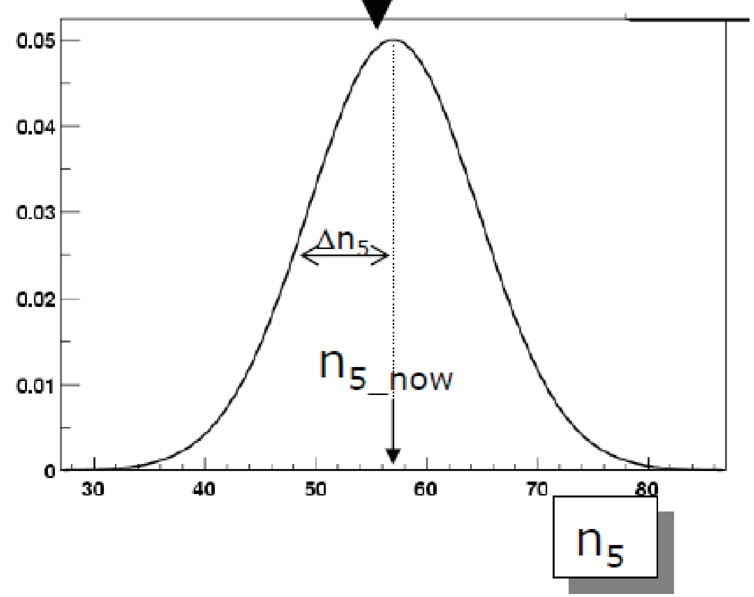
The # of entries $n_5$ in bin 5 distributed according to a Poisson distribution.

Usually: Take $n_{5\_now} = E[n_5] = \nu$

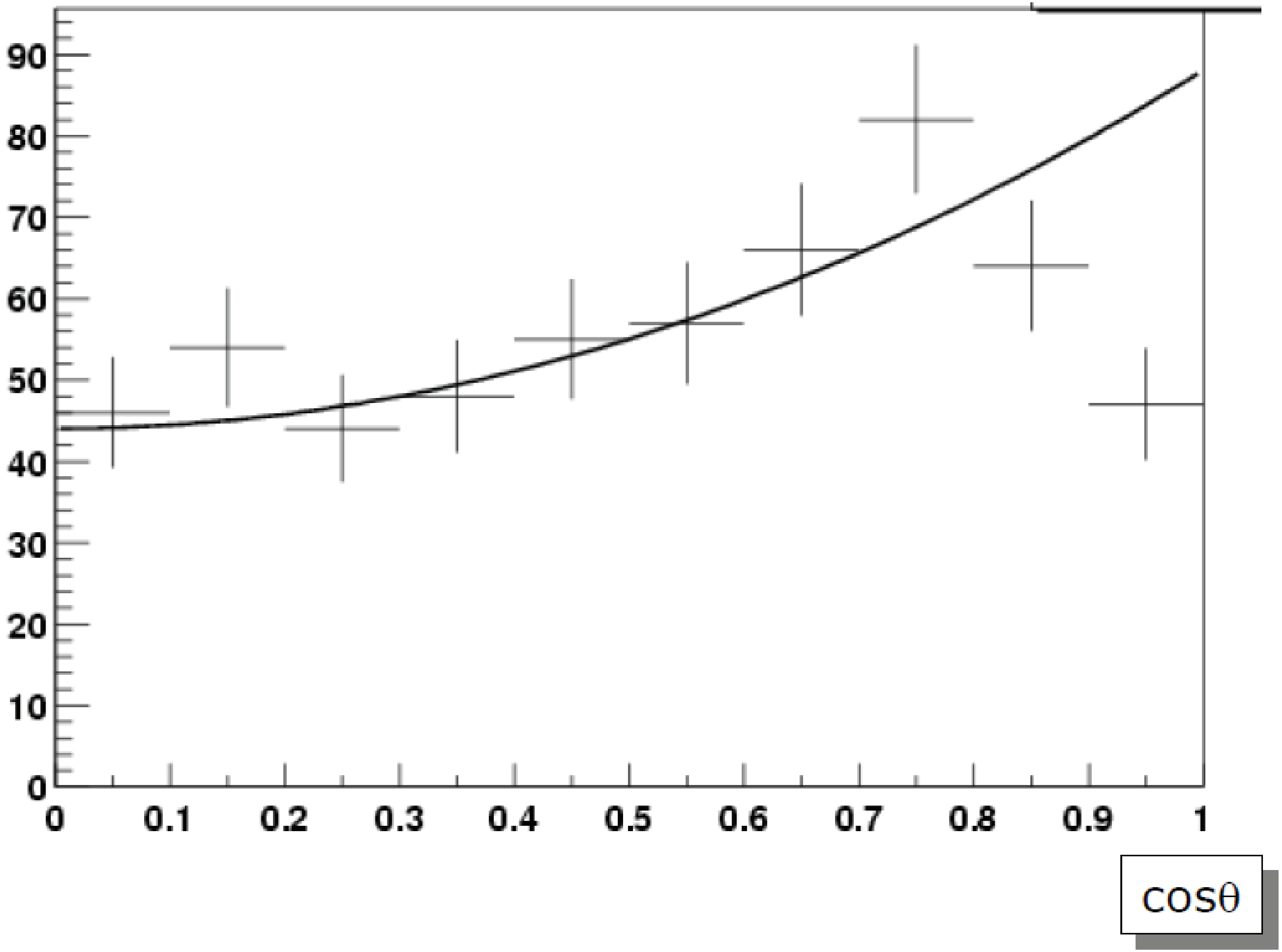
$\Rightarrow \Delta n_5 = \mathrm{sqrt}(\nu) = \mathrm{sqrt}(n_{5\_now})$

For large total number N of events: approximated by Gaussian!



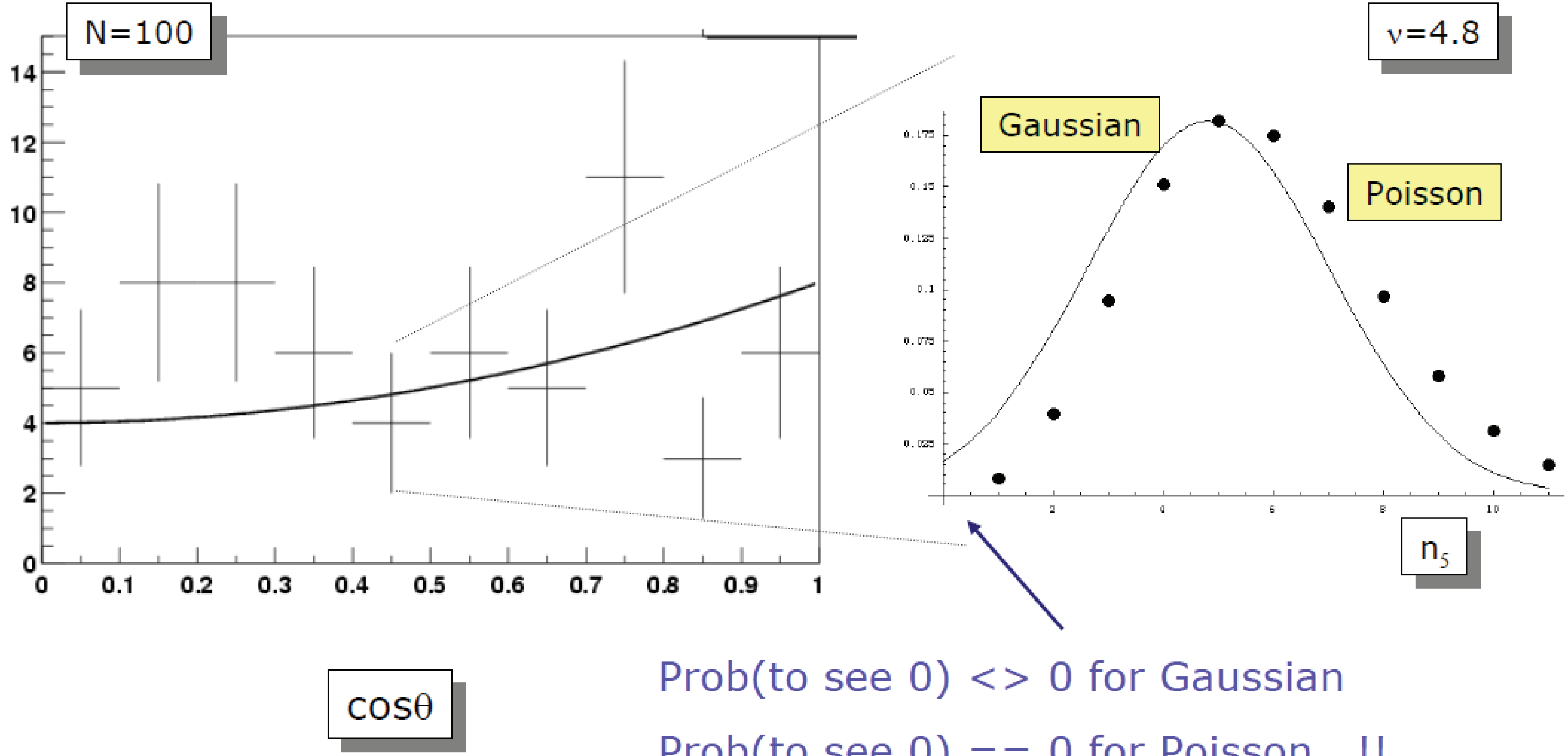



# Use errors on histogram bin values!



# Small numbers of events

Be aware that for small event numbers, Gaussian errors are wrong...




Prob(to see 0) <> 0 for Gaussian

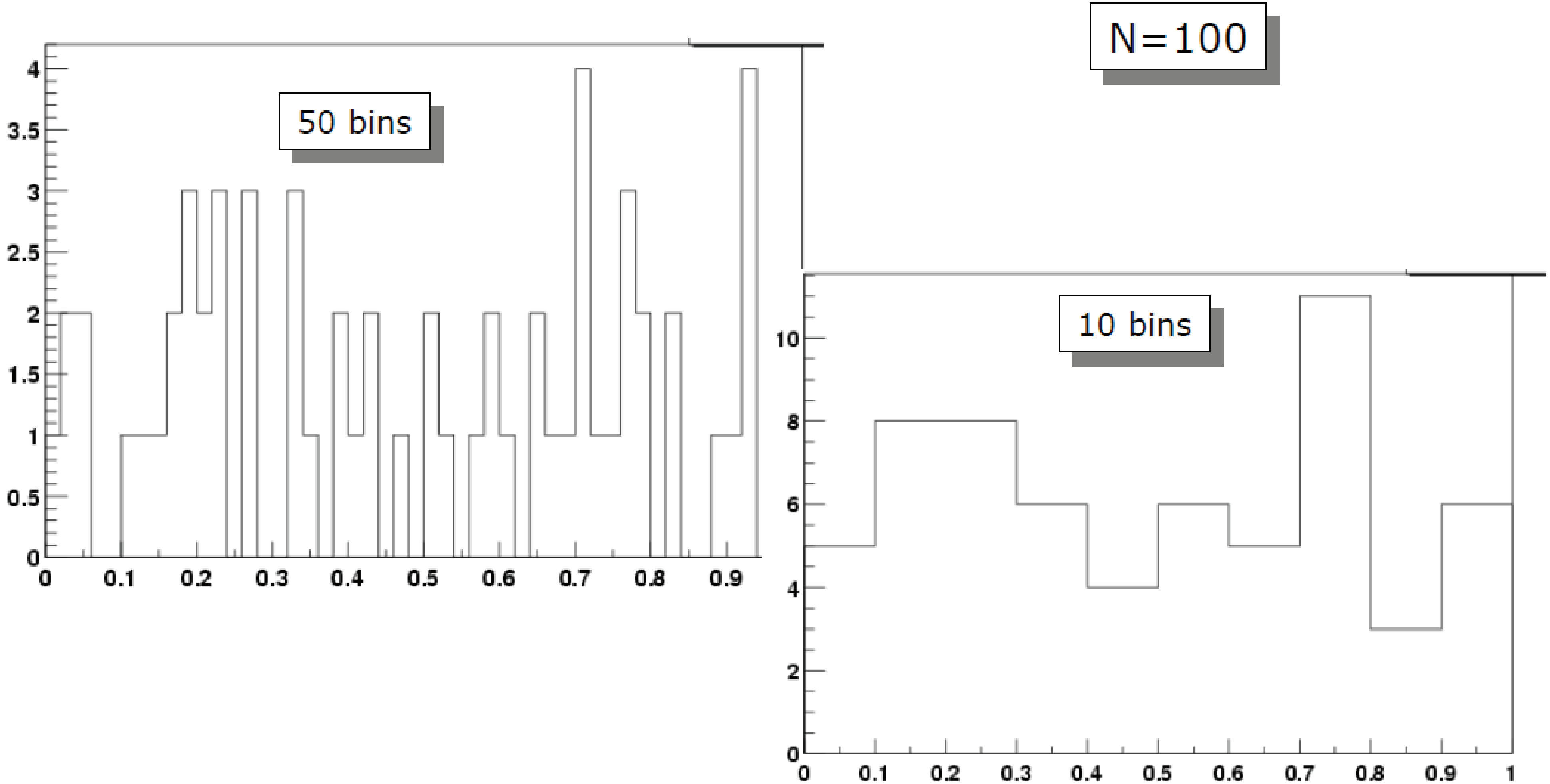
Prob(to see 0) == 0 for Poisson !!

# Histograms: Things to watch out for

- Choice of bin width
- Choice of bin range
  (underflow, overflow - important for normalisation)
- Steeply falling and quickly varying distributions

# Choice of bin width

Make sure that bins contain a reasonable number of entries

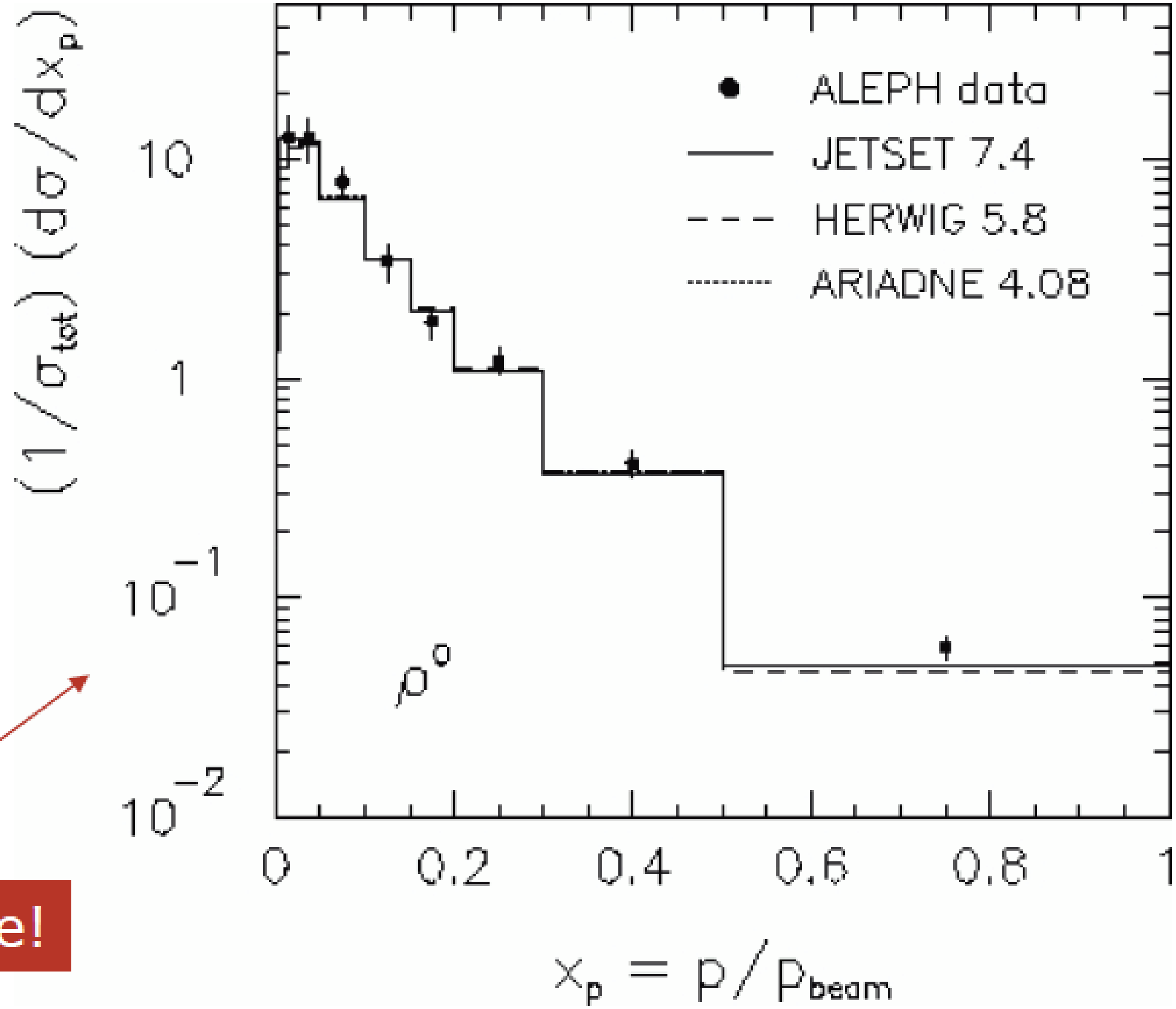


# Choice of bin width

- Take into account the experimental resolution for the variable
- Overall “statistics” (number of entries) available per bin
- Bin migration: Number of events migrating into and out of bin (due to resolution) should balance

# Choice of bin width

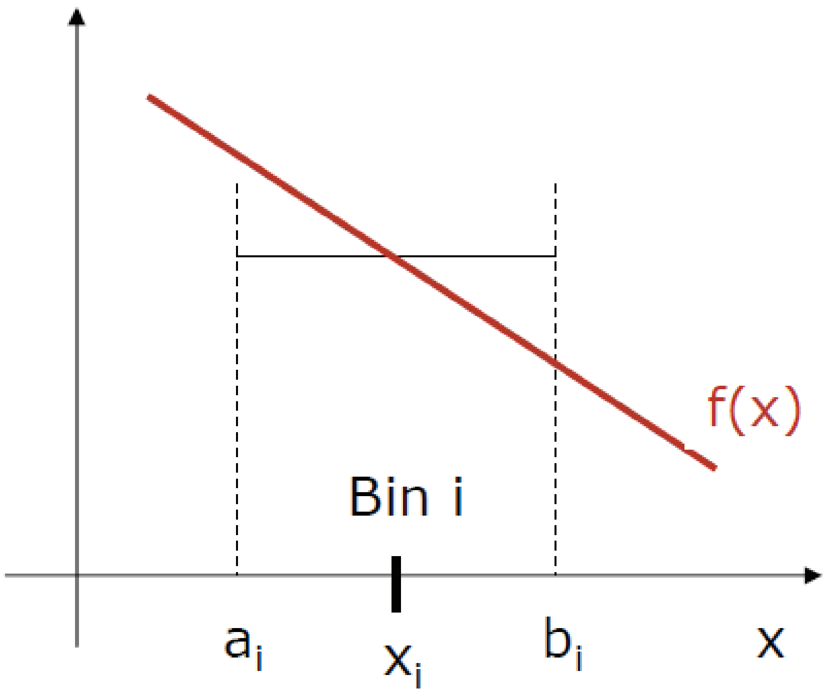
Example: Steeply falling (momentum) distribution

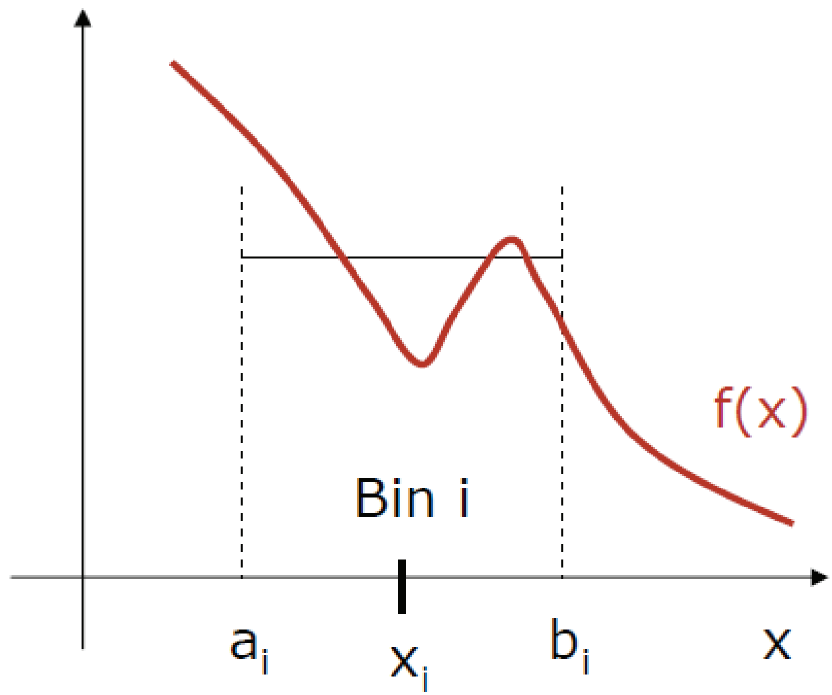



logarithmic scale!

# Comparing histograms and smooth distributions

- Watch out for very steep or quickly changing functions



# Parameter estimation and fitting

# Parameter estimation and fitting

- Set of measurements $x_i$
  (e.g. lifetimes of individual pions)

- Assumed to be distributed according to a pdf with free parameter(s)
  (e.g. an exponential distribution for a lifetime $\tau$)

- Determine an estimate of the free parameter from the data
  (fit for the lifetime $\tau$)

- Most commonly used methods:
  - Least squares
  - Maximum likelihood

# Method of least squares

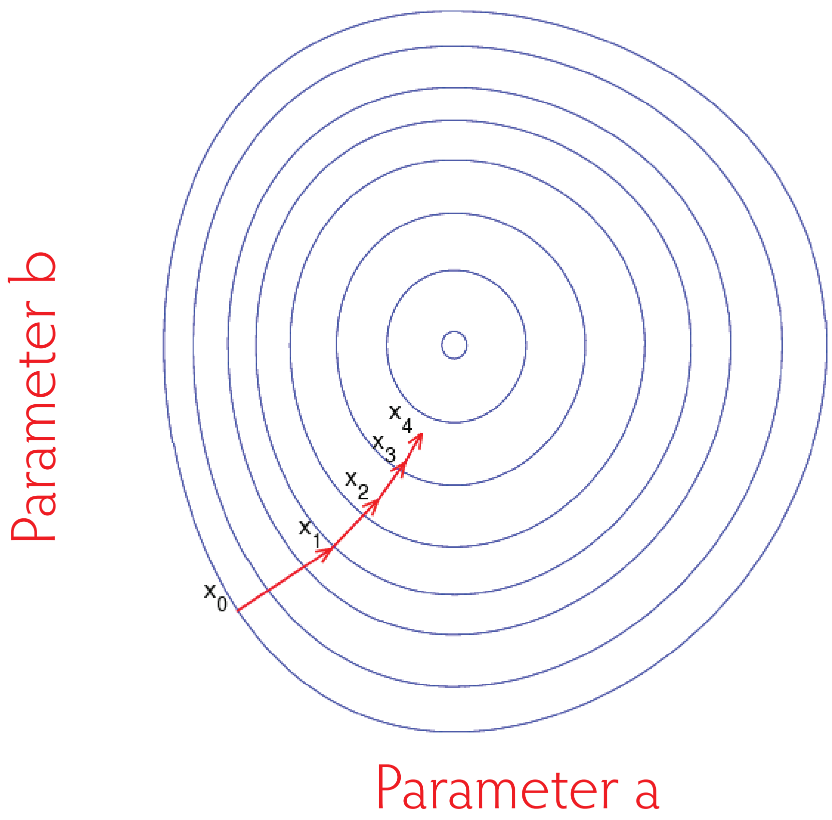
- Set of measurements ($y_i \pm \sigma_i$)
- Calculate the $\chi^2(a)$ function with parameters a, using the fit function f(x,a):

$$X^2(a) = \sum_{i=1}^{N} \frac{[y_i - f(x_i;a)]^2}{\sigma_i^2}$$

- Best estimate for a is obtained by minimizing $\chi^2(a)$
- For histograms: Bin content of bin i can be interpreted as $y_i$

# In practice

- Fitting of functions to histograms is built into data analysis packages (e.g. root, see tomorrow)
- The actual minimizing is done by a time honoured software package called MINUIT (gradient descent method)



# Least squares...

Look at goodness of fit!

- By eye! Fit function and histogram should be similar

- The $\chi^2$ is a measurement of the goodness of fit
  (for a fixed number of degrees of freedom)

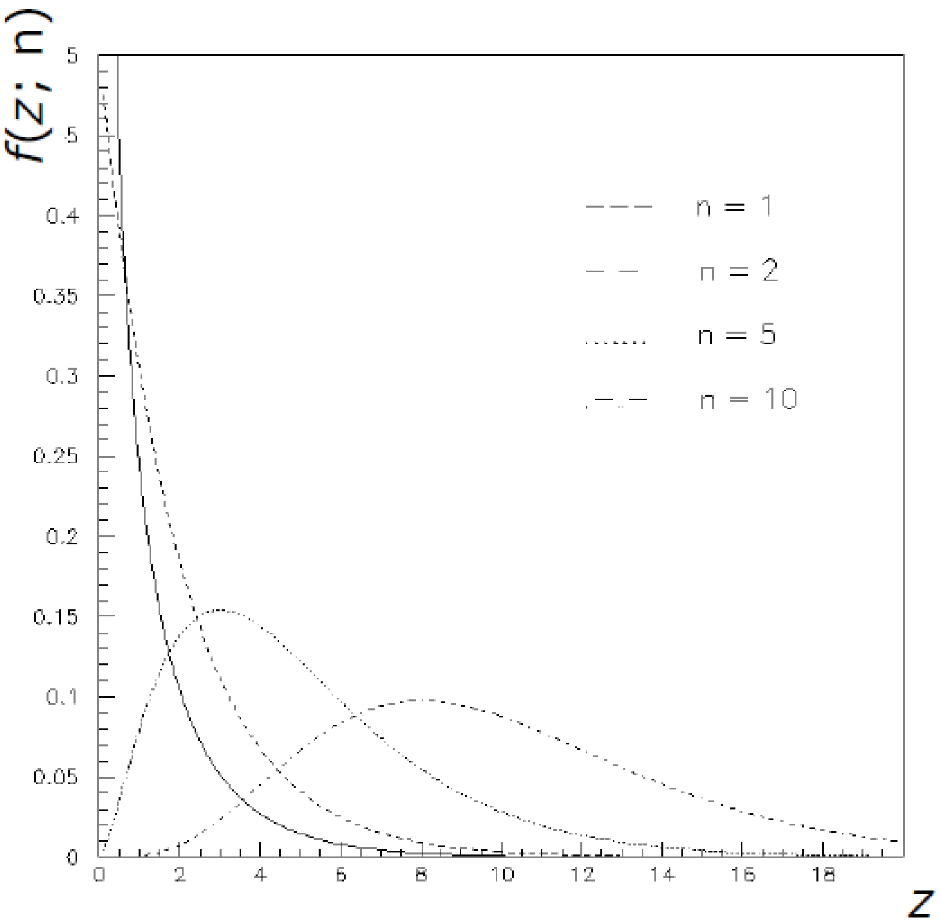
- If the data are Gaussian distributed, variances are known, the model is linear in the fit parameters, and it is the right model then:
  - $\chi^2$ sum is distributed according to the $\chi^2$ distribution
  - Expectation value =
    number of degrees of freedom =
    number of bins - number of parameters
  - Prob($\chi^2$, ndf) is flat
  - if $\chi^2$ >> ndf: Bad fit: error estimates to small, model wrong, minimization failed
  - if $\chi^2$ << ndf: Error estimates to large

# Reminder: $\chi^2$ distribution

- If $x_1 \ldots x_n$ are independent, Gaussian distributed variables with mean $\mu$ and variance $\sigma$, then

$$z = \sum_n ((x_i - \mu)/\sigma)^2$$

is distributed according to the $\chi^2$ distribution




$$f(z;\ n) = \frac{1}{2^{n/2}\Gamma(n/2)} z^{n/2-1} e^{-z/2} \quad ; \quad n = 1, 2, \ldots$$

$$E[z] = n$$

$$V[z] = 2n$$

Mean is = n =
number of degrees of freedom

# Least squares: Pro and con

## Advantages

- Easy to use (implement)
- Fast (also for huge data samples)
- Goodness of fit estimate available
- Useful general method to compare two distributions

## Disadvantages

- Information lost due to binning
- Have to be very careful with bins with few entries:
  - Need some ≥ 10 entries
  - No zeroes

  Else: Errors non-Gaussian, do not expect $\chi^2$ distribution
- Be careful if there are large bin-to-bin correlations
  (need to invert covariance matrix)

# Maximum Likelihood

- Set of measurements $x_i$

- Calculate the Likelihood function with parameters a, using the fit function f(x,a):

$$L=\prod_{i=0}^{n} f(x_i, a)$$

- Then go to the negative logarithm of the Likelihood function

$$-\log L=-\sum_{i=0}^{n} \log f(x_i, a)$$

- Minimize this function to obtain an estimate of the parameter(s) a

# Maximum likelihood: Pro and con

Advantages

- No loss of information due to binning
- Good for very uneven pdfs
- No requirements on linearity of model
- No issues with correlations if events are independent

- For $n \rightarrow \infty$: Is the best possible estimator

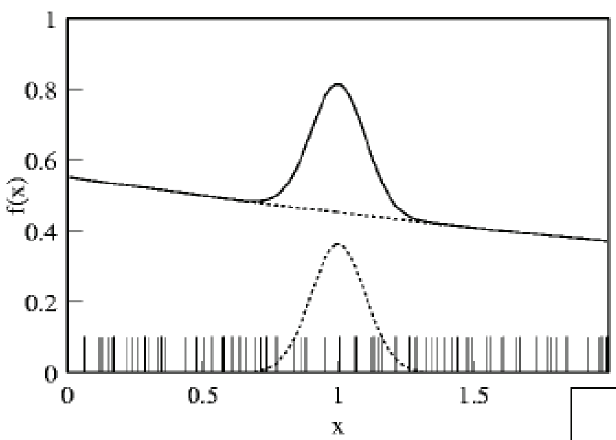
Disadvantages

- A bit more tedious to implement
- Can be slow for large data sets
- No absolute goodness of fit
- Model needs to be normalised

# Example: Signal and Background

What if the data contain contributions from different sources?

- Add different pdfs…
- Example: Search for a new resonance, after selection, data still contain some background



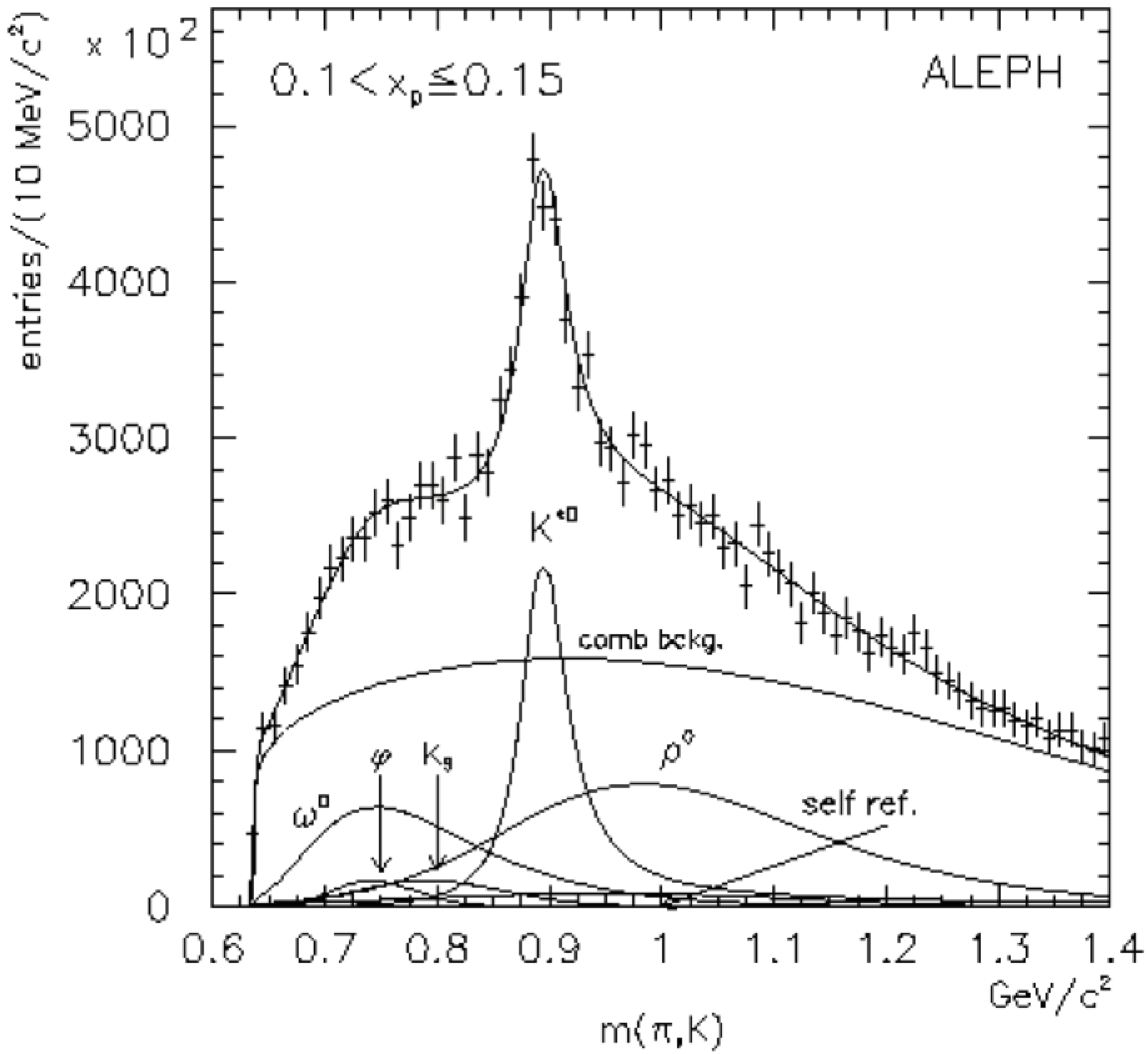

number of different contributions

$$f(x;\,\theta) = \sum_{i=1}^{m} \theta_i f_i(x)$$

can depend on further parameters

relative fractions :
a priori known (from analytical calc. or Monte Carlo), or to be fitted!

# Can have many components



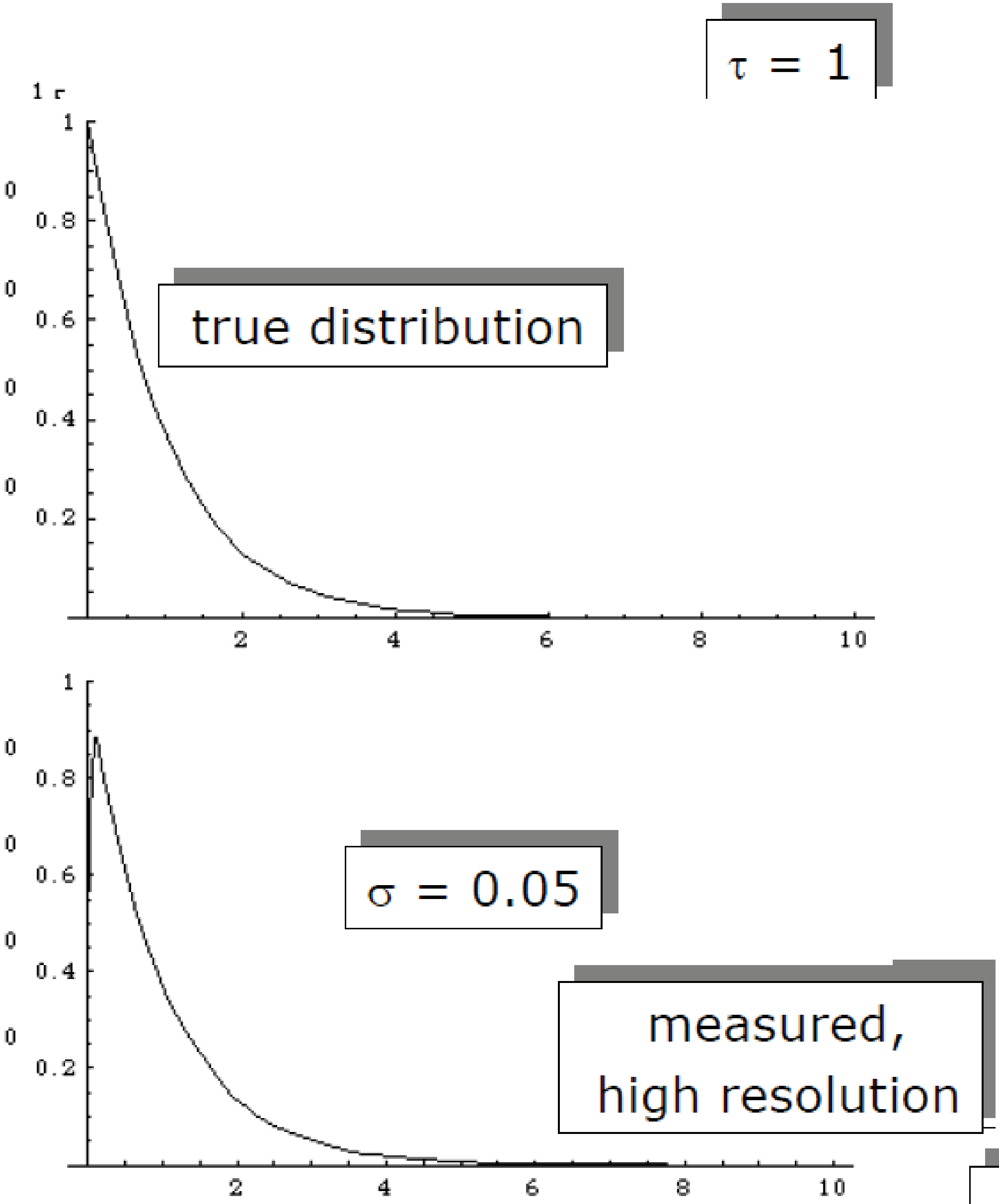
# Example: Taking into account resolution

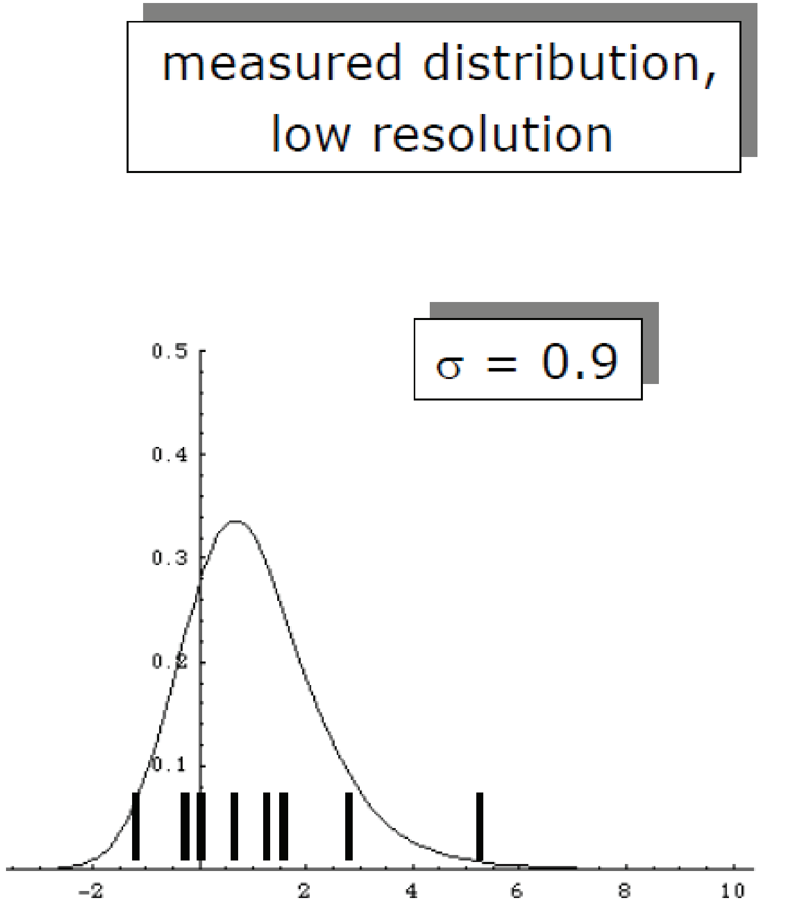
Performing a lifetime measurement with a finite time resolution

- Lifetimes distributed exponentially, with lifetime τ: f(τ,t)
- Measurements smeared with a resolution σ (assume Gaussian) around their true value R(t,t')
- Measured distribution will be a convolution of the two:

$$M(\tau;t) \equiv \int R(t,t') \cdot f(\tau;t) \cdot dt'$$

# Measurement with resolution




so: can measure negative values!

Take this into account when fitting for $\tau$

# Uncertainties (errors)

# Counting errors

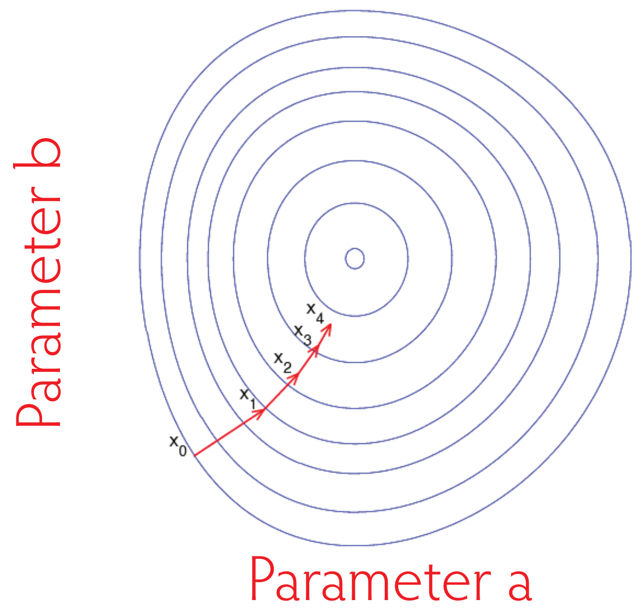
The accuracy of the measurement will be limited by the number of data events

- For N large, the statistical error goes as $\sqrt{N}$
- For N towards infinity, the relative error goes to 0

# Fit Errors

- MINUIT returns parameters and errors
- Error given by change of objective function by 1 ($\chi^2$) or 0.5 (log LH)
- MINUIT normally estimates error from gradient at minimum
- Calling HESSE after MINUIT also gives you correlations (the error matrix)
- MINOS will actually scan the parameters and return asymmetric errors

- Fit errors DO NOT tell you about the goodness of fit
  (only about the size of your data sample)



# Statistical errors

The accuracy of the measurement will be limited by the number of data events

- For N large, the statistical error goes as $\sqrt{N}$
- For N towards infinity, the relative error goes to 0

But is the large N measurement really arbitrarily precise?

# Systematic errors

No, the measurement can still be systematically off

- Clock running slow
- Calorimeter not perfectly calibrated
- Cable delays not properly accounted for
- Fitting an inadequate model
- etc.

These errors lead to systematic uncertainties

- Description of how well we understand the measurement

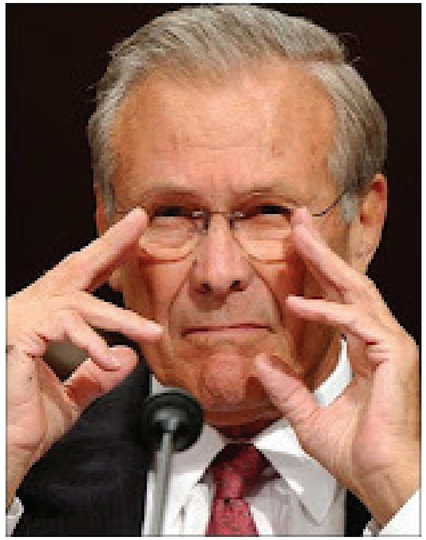
# Systematic errors

“ [T]here are known knowns; there are things we know that we know.

There are known unknowns; that is to say there are things that, we now know we don't know.

But there are also unknown unknowns – there are things we do not know, we don't know. ”

—United States Secretary of Defense,
Donald Rumsfeld



# Systematic errors

Determining statistical errors is a science

Estimating systematic errors is an art

# Systematic errors

Estimating systematic errors is a very important part of the analysis

- What assumptions went into the measurement?
- How well do you understand these assumptions?
- Can you make auxiliary measurements to test assumptions/obtain calibrations?

  e.g. use a beam of particles of known energy to calibrate a calorimeter

- You can never be totally sure that you have taken into account every single possible effect
- Think about systematics before starting the analysis

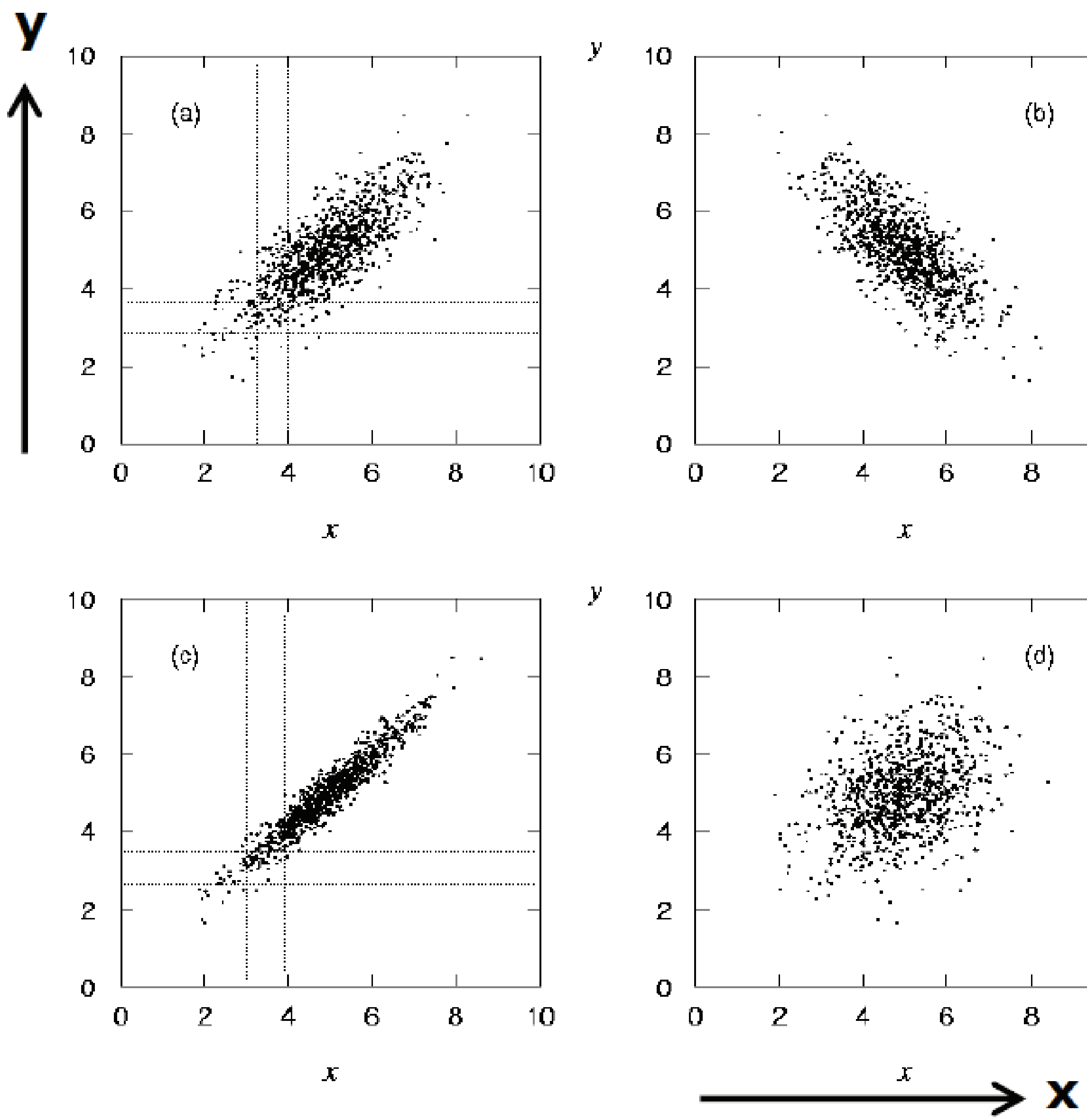
# Correlations

# More than one variable

- Let f(x, y) be the joint pdf to observe
  x in [x, x + dx]
  y in [y, y + dy]

- Useful tool here: 2D-histograms, often drawn as scatterplots

- f(x,y) = density of points = #entries



# Covariance/correlations

- Let f(x, y) be the joint pdf

- If the variables are independent, then x and y are uncorrelated:
  The joint pdf factorizes: f(x, y) = g(x) h(y)

- For correlated variables, define the covariance between two variables x, y:
  cov(x, y) = V(x, y)

$$\mathrm{cov}(x,y) = \langle (x - \langle x \rangle)\cdot(y - \langle y \rangle) \rangle$$
$$= \langle xy \rangle - \langle x \rangle \langle y \rangle$$

- Properties:
  - cov(x, x) = V(x)
  - cov(x, y) is translation invariant (shift origin) and has units
  - V(x + y) = V(x) + V(y) + 2 cov(x, y)

# Covariance/correlations

The covariance can be represented by a matrix

$$V(x,y) = \begin{pmatrix} \sigma_x^2 & V[xy] \\ V[yx] & \sigma_y^2 \end{pmatrix}$$

$$V[xy] = E((x-\mu_x)(y-\mu_y)) \equiv E[xy] - \mu_x \mu_y$$

$$E[xy] = \int_{y\min}^{y\max} \int_{x\min}^{x\max} x' \cdot y' \cdot f(x', y') \cdot dx' \cdot dy'$$

we used here true values $\mu_x$ and $\mu_y$ instead of <x>, <y>

- V(x, y) is often called the error matrix;
  the diagonal elements are just the variances

# Correlation coefficient

Define correlation coefficient ρ

- ρ ranges between -1 and +1
- If the variables are uncorrelated, ρ=0
- The opposite is not true

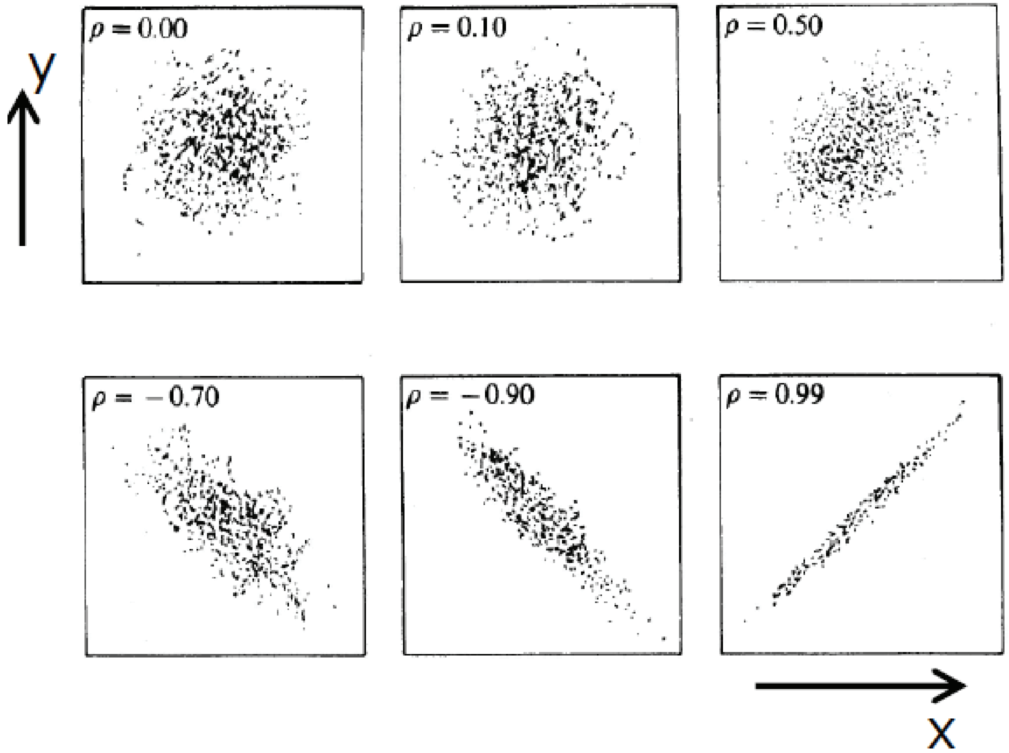
$$\rho_{xy} = \frac{\mathrm{cov}(x,y)}{\sqrt{V(x)\cdot V(y)}} = \frac{V(x,y)}{\sigma_x \sigma_y}$$

An estimate for ρ is $r_{xy}$, taken from the sample variance $s_{xy}$:

$$r_{xy} = \frac{s_{xy}}{\sigma_x \sigma_y}$$

$$s_{xy} = \frac{1}{n-1}\sum_i (x_i - \langle x \rangle)(y_i - \langle y \rangle)$$

$$\sigma_x = \sqrt{V(x)}$$



# Correlation coefficient: Questions

An estimate for ρ is $r_{xy}$, taken from the sample variance $s_{xy}$:

$$r_{xy} = \frac{s_{xy}}{\sigma_x \sigma_y}$$

$$\rho_{xy} = \frac{\mathrm{cov}(x, y)}{\sqrt{V(x) \cdot V(y)}} = \frac{V(x, y)}{\sigma_x \sigma_y}$$

$$s_{xy} = \frac{1}{n-1} \sum_i (x_i - \langle x \rangle)(y_i - \langle y \rangle)$$

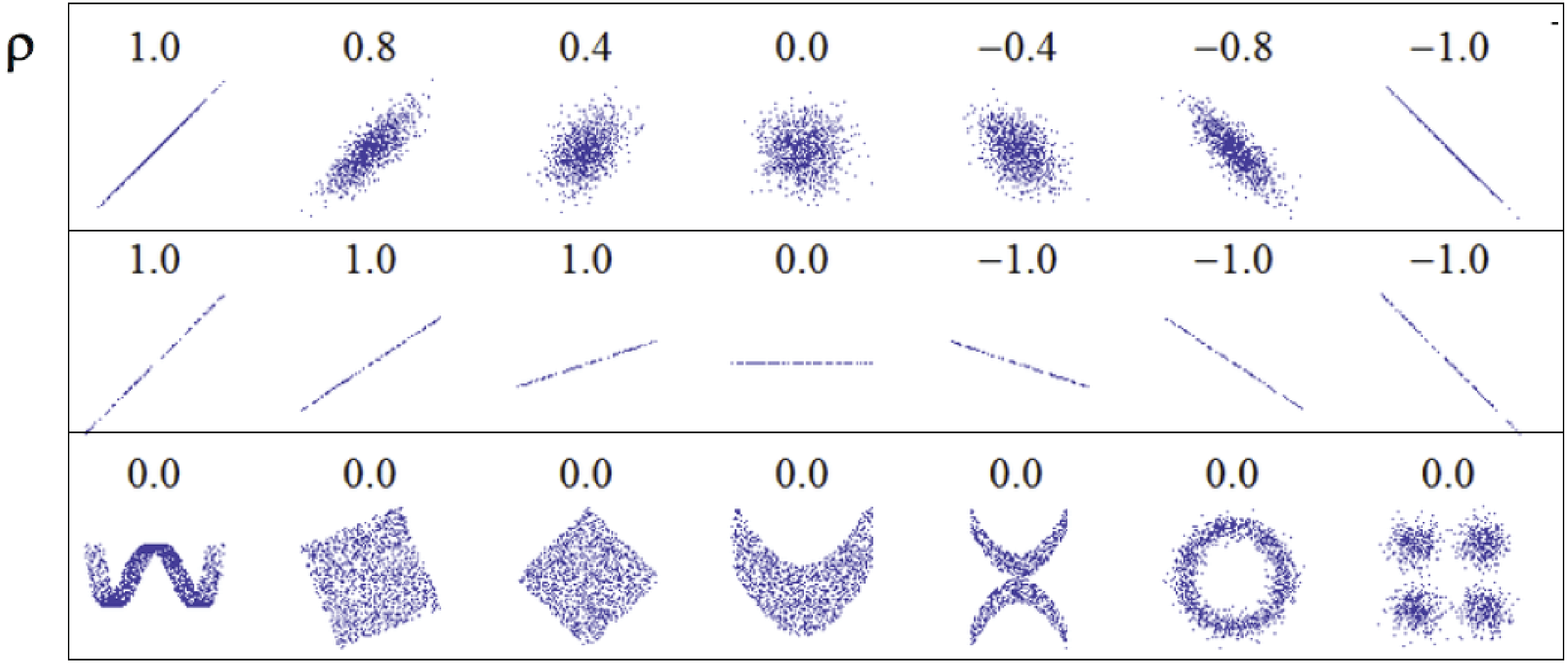
$$\sigma_x = \sqrt{\mathrm{V}(\mathrm{x})}$$

What is the correlation coefficient for (x, y) on a horizontal line? A vertical line?

What is the correlation coefficient for (x, y) on a circle?

# Overview

- Correlation coefficient reflects the direction of a linear relationship
- It does not reflect the slope
- It does not reflect many properties of nonlinear relationships
  ρ = 0 does not imply no correlation



# Error propagation

- For uncorrelated variables:

$$\sigma_f^2 = \sum_{i=1}^{n} \left( \frac{\partial f}{\partial x_i} \right)^2 \cdot \sigma_{x_i}^2$$

- If they are correlated, take this into account:

$$\sigma_f^2 = \sum_{i=1}^{n} \left( \frac{\partial f}{\partial x_i} \right)^2 \cdot \sigma_{x_i}^2 + \sum_{i=1}^{n} \sum_{j \neq i}^{n} \left( \frac{\partial f}{\partial x_i} \frac{\partial f}{\partial x_j} \right) \cdot \mathrm{cov}(x_i, x_j)$$