Karlsruhe Institute of Technology

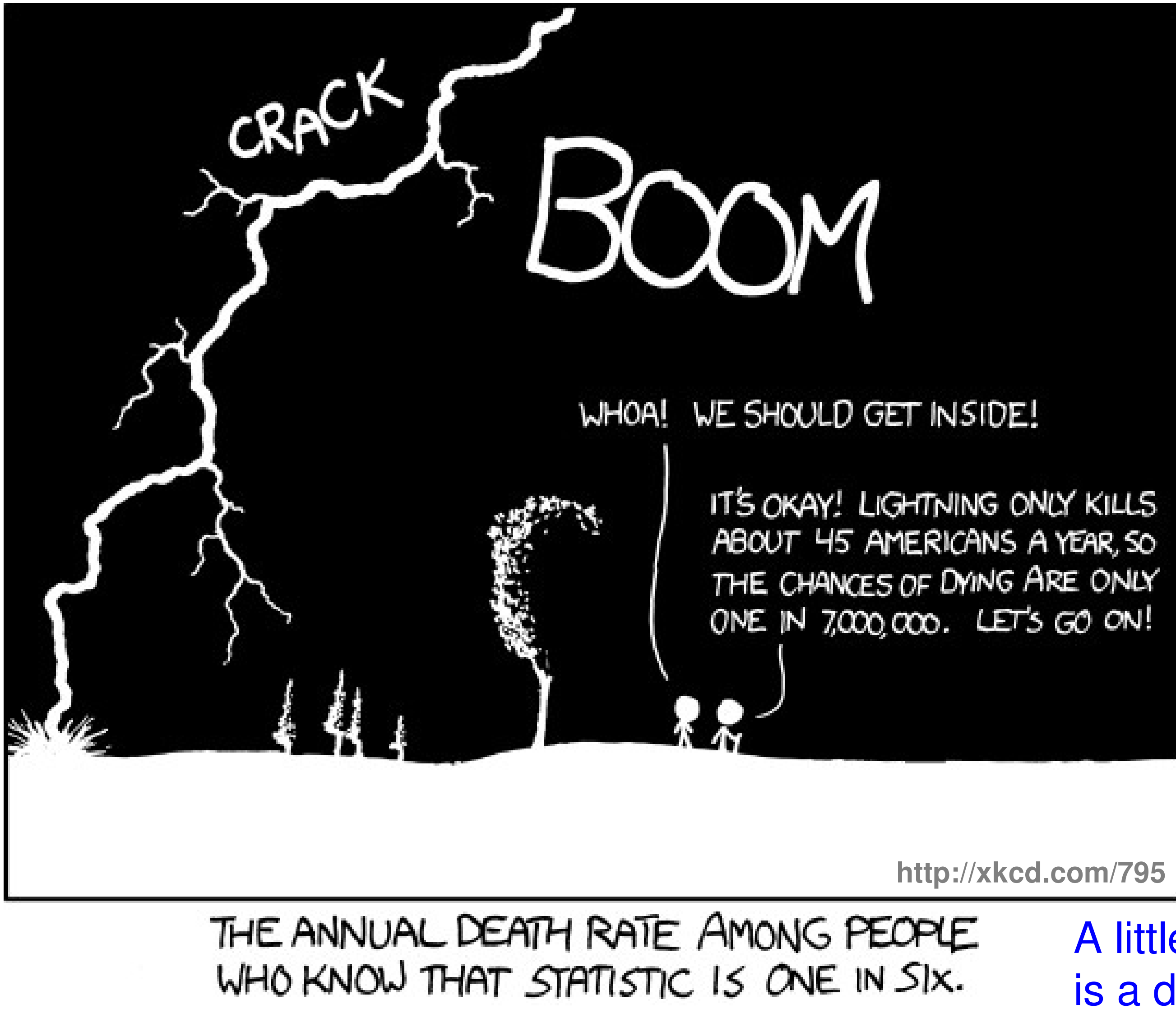
Statistische Methoden der Datenanalyse

Hans Dembinski
IEKP, KIT Karlsruhe

Idea of this lecture

- Show common statistical tools and best practice methods
- Explain basics and foundations
- Use lots of examples (be practical, but simple)
- See textbooks for completeness, details, and proofs

Humor



A little knowledge
is a dangerous thing...

Lecture summary

- Thursday
  - Probability
  - Model fitting
  - Confidence intervals
- Friday
  - Confidence limits
  - Monte-Carlo and resampling methods
  - Testing hypotheses
- Saturday
  - Probability density estimation
  - Multivariate classification
  - Optional: Artificial neural networks

Literature

- G. Cowan, *Statistical data analysis*, Claredon Press (1998)
- F. James, *Statistical Methods in Experimental Physics* – 2nd edition, World Scientific (2006)
- B. Efron and R. Tibshirani, *An introduction to the bootstrap*, Chapman and Hall (1993)
- V. Blobel and E. Lohrmann, *Statistische und numerische Methoden der Datenanalyse*, Teubner Verlag (1998)
- A. J. Izenmann, *Modern Multivariate Statistical Techniques*, Springer (2008)
- TMVA Workshop @ CERN, January 2011 – http://indico.cern.ch/event/tmva_workshop
- Davison and Hinkley, *Bootstrap methods and their applications*, Cambridge University Press (1997)
- Press, Teukolsky, Vettering, Flannery, *Numerical Recipes* – 3rd edition, Cambridge University Press (2007)

Useful software

Python + numpy + scipy + matplotlib

www.python.org www.scipy.org www.numpy.org matplotlib.sourceforge.net

ROOT (in particular RooFit, RooStats)

root.cern.ch/drupal

R (main tool of statisticians)

www.r-project.org

TMVA

tmva.sourceforge.net

Topics for today

- Probability
  - Bayesian and Frequentist views
  - Bayes theorem
  - Probability distributions and probability density functions
- Model fitting
  - Maximum-likelihood method
  - (Linear) least-squares method
- Calculation and interpretation of fit uncertainties

Probability

Probability

Frequentist view

P = frequency of outcome from a (in principle) repeatable process

Objective statements

Confidence regions based on **coverage**

Bayesian view

P = degree of belief (betting odds!)

Allows one to calculate *P* of non-repeatable events, e.g. “probability” of a theory being correct

No objective statements

Results depend on *prior beliefs*

Can handle *systematic uncertainties*

Calculus for probabilities

Both Bayesian and Frequentist probabilities obey the *Kolmogorov axioms*

Let's regard a set of exclusive events X_i with probability $P(X_i)$ of occurrence of X_i

$a)\ P(X_i) \geq 0 \text{ for all } i$ — probabilities cannot be negative

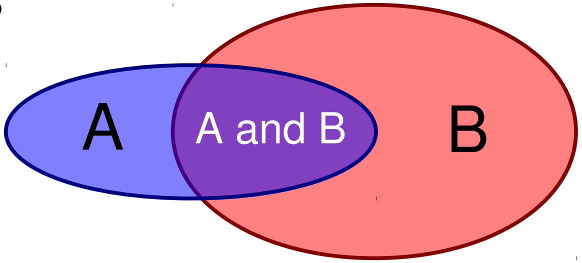
$b)\ P(X_i \text{ or } X_j) = P(X_i) + P(X_j)$ — probabilities of mutually exclusive events add up

$c)\ \sum_i P(X_i) = 1$ — probabilities of all mutually exclusive events add up to one

More general rules follow for non-exclusive events

$$P(A \text{ or } B) = P(A) + P(B) - P(A \text{ and } B)$$

$$P(A \text{ and } B) = P(A|B)P(B) = P(B|A)P(A)$$



A and B are independent if $P(A|B) = P(A)$, then $P(A \text{ and } B) = P(A)P(B)$

Bayes theorem
(Bayesian **and** Frequentist)

$$P(A_i|B) = \frac{P(B|A_i)\,P(A_i)}{P(B)} = \frac{P(B|A_i)\,P(A_i)}{\sum_i P(B|A_i)\,P(A_i)}$$

Bayesian use of Bayes theorem

After looking at LHC data, should I believe in the Higgs? Use Bayes theorem:

$$P(\mathrm{Higgs}|\mathrm{data}) = \frac{P(\mathrm{data}|\mathrm{Higgs})\, P(\mathrm{Higgs})}{P(\mathrm{data}|\mathrm{Higgs})\, P(\mathrm{Higgs}) + P(\mathrm{data}|\mathrm{no\,Higgs})\, P(\mathrm{no\,Higgs})}$$

What is my prior belief in the Higgs?
I don't know.

$$P(\mathrm{Higgs}) = P(\mathrm{no\,Higgs}) = 0.5$$

Uninformative prior

Use of Bayes theorem with *uninformative priors* is the closest to objective inference that Bayesian methodology has to offer

a) $P(\mathrm{data}|\mathrm{Higgs}) = 0.6 \qquad P(\mathrm{data}|\mathrm{no\,Higgs}) = 0.1 \quad \Rightarrow \quad P(\mathrm{Higgs}|\mathrm{data}) = 0.75$

Odds to explain data with/without the Higgs 6 to 1, still the Higgs is not a sure bet

b) $P(\mathrm{data}|\mathrm{Higgs}) = 0.8 \qquad P(\mathrm{data}|\mathrm{no\,Higgs}) = 0.8 \quad \Rightarrow \quad P(\mathrm{Higgs}|\mathrm{data}) = 0.5$

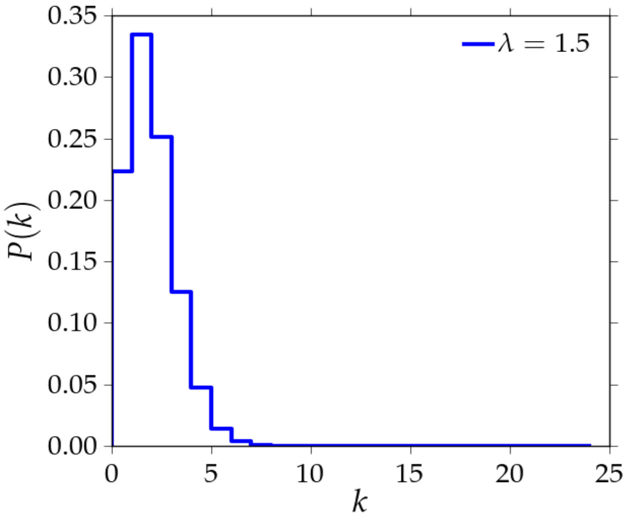
Data did not allow to discriminate between the hypotheses, no update of my belief

Probability distributions

Discrete outcomes (e.g. event/particle counts)

Expectation $E[k] = \sum_i k_i P(k_i)$ Variance $V[k] = \sum_i k_i^2 P(k_i) - \left(\sum_i k_i P(k_i)\right)^2$

Poisson

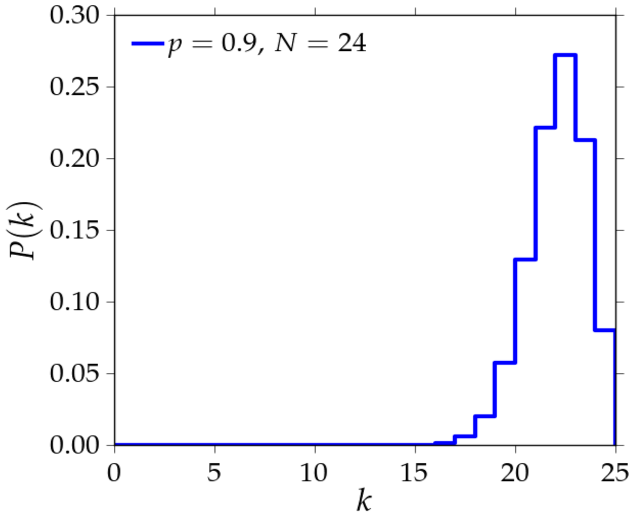


$$P(k|\lambda) = \frac{e^{-\lambda}\,\lambda^k}{k!}$$

$$E[k] = \lambda \qquad V[k] = \lambda$$

Count of events from a source

Binomial



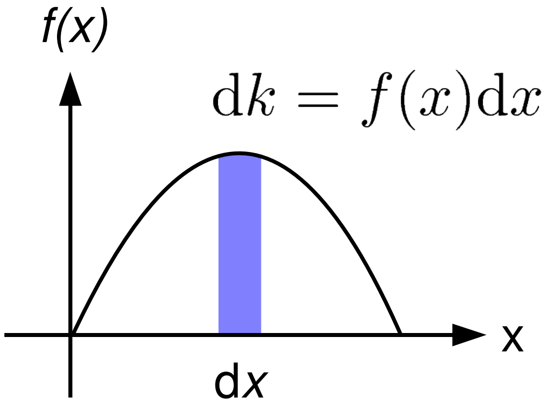
$$P(k|p,N) = \binom{N}{k} p^k\,(1-p)^{N-k}$$

$$E[k] = Np \qquad V[k] = Np(1-p)$$

Selection of k events out of N events

Probability distributions

Continuous outcomes (e.g. energy deposited in a detector)



$$\mathrm{d}k = f(x)\mathrm{d}x$$

$$E[x] = \int \mathrm{d}x\, x f(x) \qquad E[g(x)] = \int \mathrm{d}x\, g(x) f(x)$$

Linearity $E[a\,x + b\,y] = a\,E[x] + b\,E[y]$

In general for non-linear *g(x)* $E[g(x)] \neq g(E[x])$

$$V[x] = E[x^2] - E[x]^2$$

$$V[a\,x + b\,y] = a^2\,V[x] + b^2\,V[y] + 2ab\,\mathrm{cov}[x, y]$$

Multivariate case

$$\mathrm{d}k = f(\vec{x})\,\mathrm{d}\vec{x} = f(\vec{x})\,\mathrm{d}x_0 \cdots \mathrm{d}x_n$$

$$E[g(\vec{x})] = \int \mathrm{d}\vec{x}\, g(\vec{x}) f(\vec{x}) = \int \mathrm{d}x_0 \cdots \mathrm{d}x_n g(\vec{x}) f(\vec{x})$$

Covariance matrix $\mathrm{cov}[x_i, x_j] = E[x_i\,x_j] - E[x_i]\,E[x_j] \qquad \mathrm{cov}[x_i, x_i] = V[x_i]$

Correlation $\mathrm{corr}[x_i, x_j] = \dfrac{\mathrm{cov}[x_i, x_j]}{\sigma[x_i]\,\sigma[x_j]} \qquad -1 \leq \mathrm{corr}[x_i, x_j] \leq 1$

Probability distributions

Continuous outcomes (e.g. energy deposited in a detector)

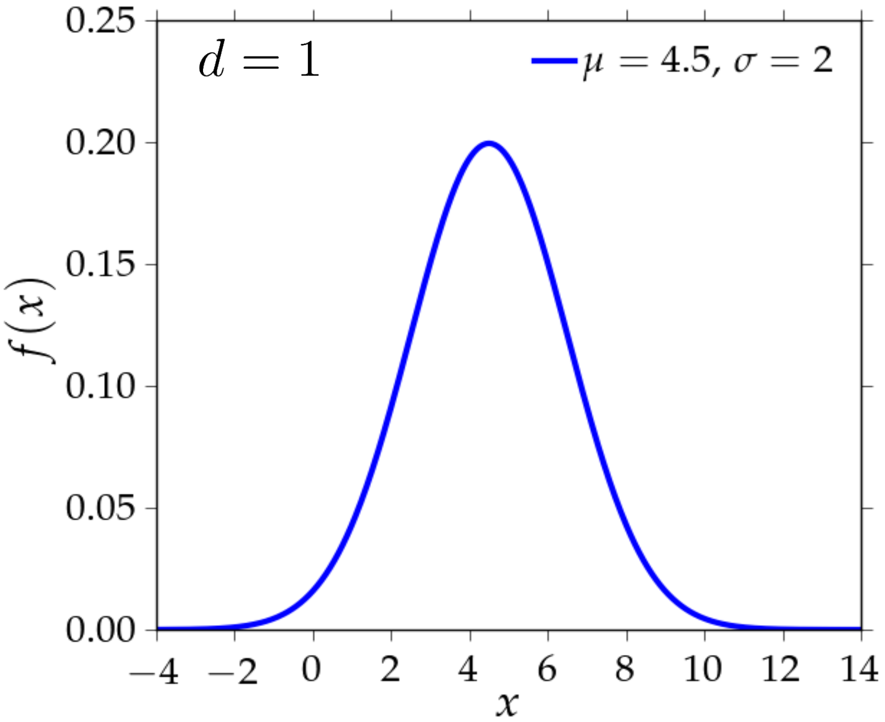
Multivariate normal (Gaussian)

$$f(\vec{x}|\vec{\mu}, V) = \frac{1}{\sqrt{2\pi}^d |V|} \exp\left(-\frac{1}{2}(\vec{x}-\vec{\mu})^T V^{-1} (\vec{x}-\vec{\mu})\right)$$

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

$$\mathrm{cov}[x_i, x_j] = V_{ij}$$

Limit of many random fluctuations added up



Probability distributions

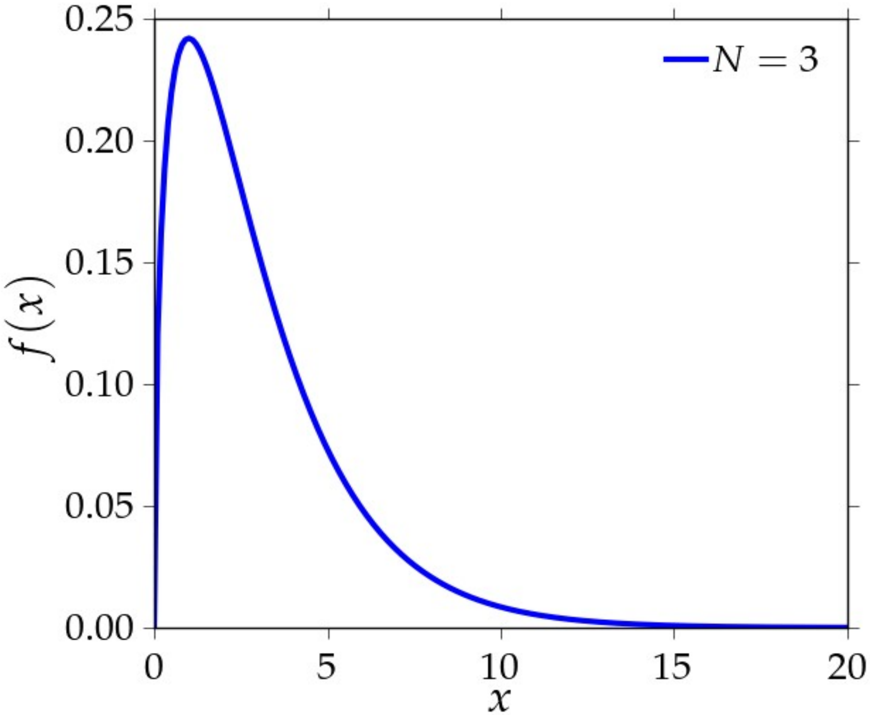
Continuous outcomes (e.g. energy deposited in a detector)

Chi-square χ^2

$$f(x) = \frac{\frac{1}{2}\left(\frac{x}{2}\right)^{N/2-1} e^{-x/2}}{\Gamma\left(\frac{N}{2}\right)}$$

$$E[x] = N$$

$$V[x] = 2N$$



Sum of N normal distributed variables with $\mu = 0$, $\sigma = 1$

Some words about correlation

Example A: x_0 and x_1 from normal distribution with $\mu,\ \sigma$

Variance of average $\bar{x} = \frac{1}{2}(x_0 + x_1)$

$$V[\frac{1}{2}(x_0 + x_1)] = \frac{1}{4}(\sigma^2 + \sigma^2) + \frac{1}{2}\underbrace{\mathrm{cov}[x_0, x_1]}_{\rho\sigma^2} = \frac{1}{2}\sigma^2(1 + \rho)$$

$\rho = 0 \Rightarrow V[\bar{x}] = \frac{1}{2}\sigma^2$ Variance decreases $\propto 1/N$

$\rho = 1 \Rightarrow V[\bar{x}] = \sigma^2$ No information gained

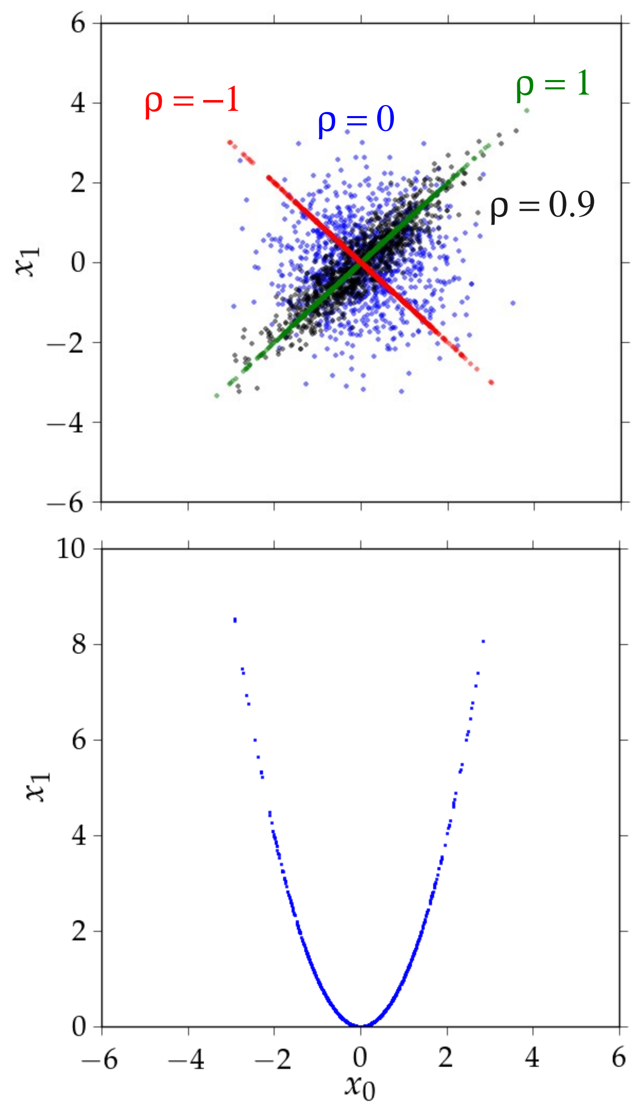
$\rho = -1 \Rightarrow V[\bar{x}] = 0$ No randomness

Independence of x_i and x_j $\not\Leftarrow$ / $\Rightarrow$ $\mathrm{cov}[x_i, x_j] = 0$

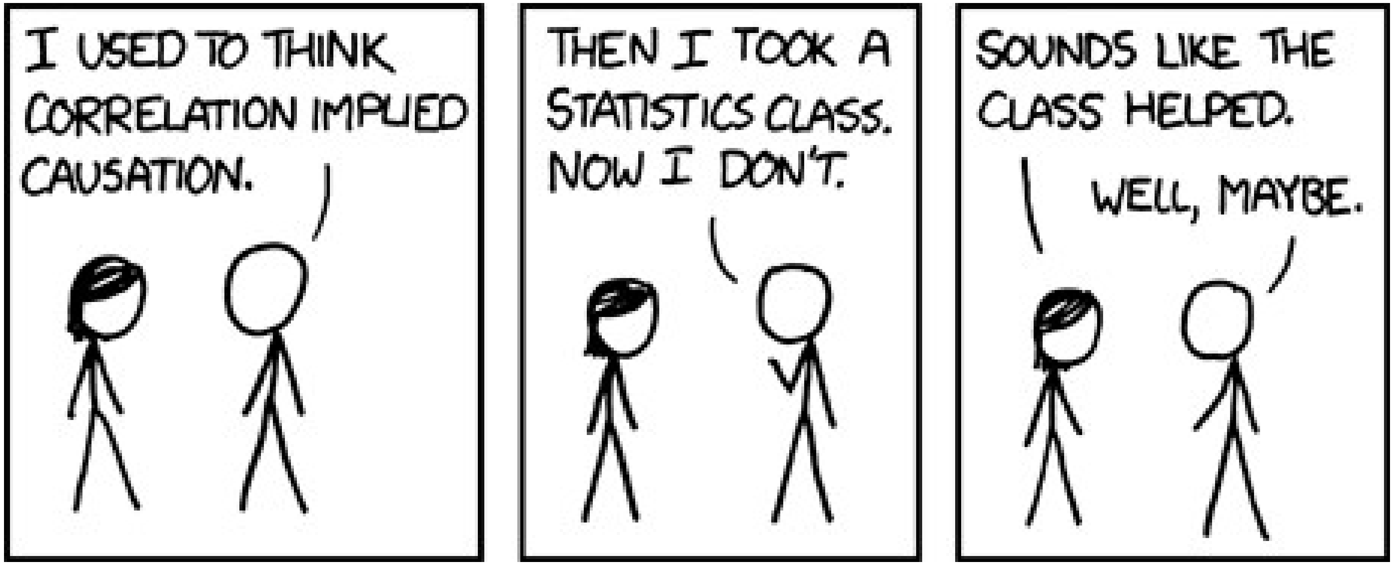
Example B: x_0 from normal distribution with $\mu = 0$, $x_1 = x_0^2$

$$\mathrm{cov}[x_0, x_1] = E[x_0\, x_1] - E[x_0]E[x_1]$$

$$= E[x_0^3] - E[x_0]E[x_0^2] = 0$$



Humor



http://xkcd.com/552

Change of variables

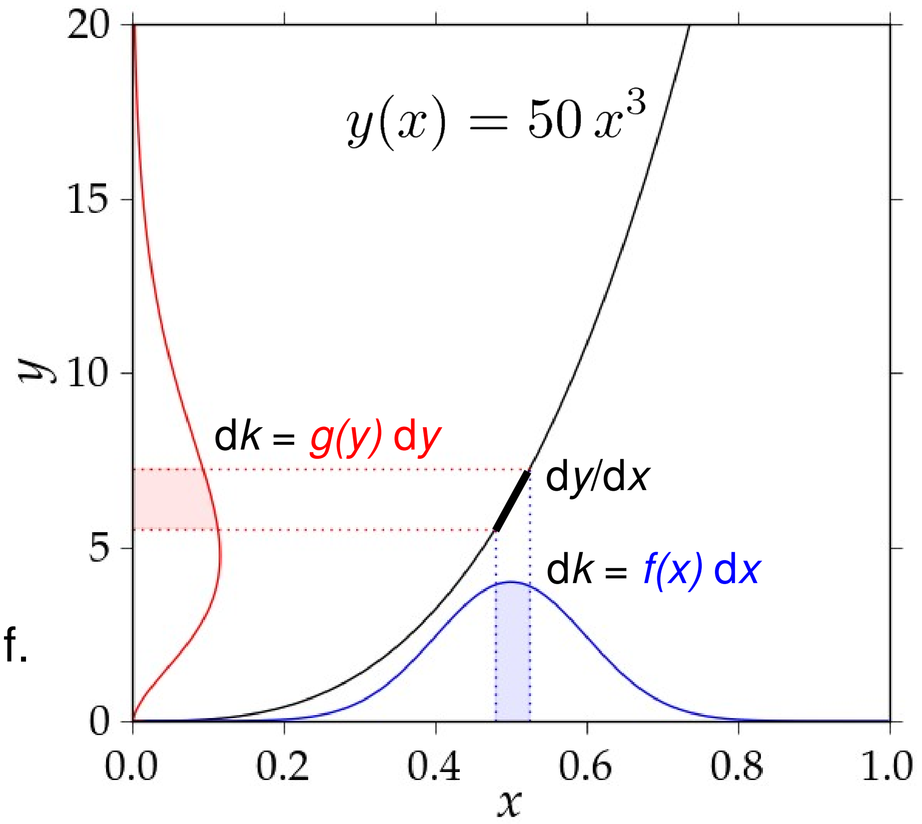
Choice of random variable of continuous distribution is usually not unique
How to transform $x \to y$?

$$\mathrm{d}k = f(x)\,\mathrm{d}x = g(y)\,\mathrm{d}y$$

$$g(y) = f(x)\left|\frac{\mathrm{d}y}{\mathrm{d}x}\right|^{-1}$$

$$g(\vec{y}) = f(\vec{x})\left|\frac{\partial\vec{y}}{\partial\vec{x}}\right|^{-1}$$

determinant of Jacobian matrix



Special case

$$y = \int_{-\infty}^{x} \mathrm{d}x'\, f(x') = F(x) \text{ c.d.f.}$$

$$g(y) = f(x)\frac{1}{f(x)} = 1 \quad \text{flat}$$

useful to judge by eye whether random variable *x* follows *f(x)*

Model fitting

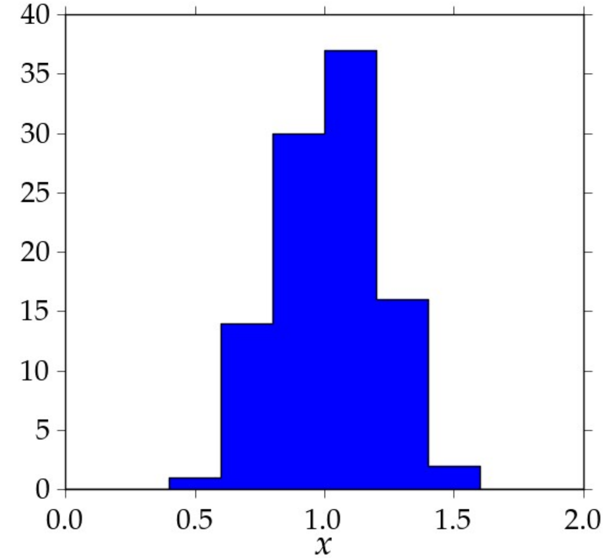
Model fitting

Unbinned data x_i or histogram $\bar{x}_i, k_i$

fitting ↑ Optimal parameters in light of data?
Uncertainty due to limited sample?

Model or empirical parametrization
with free parameters $f(x|p_1, \ldots, p_n)$

Example normal distribution
100 data points
10 bins



Maximum-likelihood method

Most general and most powerful method

Solution may depend on initial guess

Least-squares method

Good numerical properties but usually an approximation

Solution may depend on initial guess

Linear least-squares method

Fast unique solution *independent* of initial guess

use solution as starting point for full ML

Maximum-likelihood method

Idea: Model should maximize joint probability of all data points = **likelihood**

$$L(p_1, \ldots, p_n) = L(\vec{p}) = \prod_i P_i(\vec{p})$$

depends only on the model parameters $\vec{p}$

If the x_i are direct samples of a p.d.f. *f(x)*, this can be simplified

$$L(\vec{p}) = \prod_i P(x_i|\vec{p}) = \prod_i \int_{x_i}^{x_i+\Delta x_i} \mathrm{d}x\, f(x|\vec{p}) \xrightarrow[\Delta x_i \to 0]{} \prod_i f(x_i|\vec{p})\, \Delta x_i$$

we can choose the intervals arbitrarily small

Sums are easier to handle so maximize ln*L* instead of *L* (logarithm is monotonic)

$$\boxed{\ln L(\vec{p}) = \sum_i \ln P(x_i|\vec{p})} = \sum_i \ln f(x_i|\vec{p}) + \underbrace{\sum_i \Delta x_i} \equiv \boxed{\sum_i \ln f(x_i|\vec{p})}$$

constant with respect to $\vec{p}$!

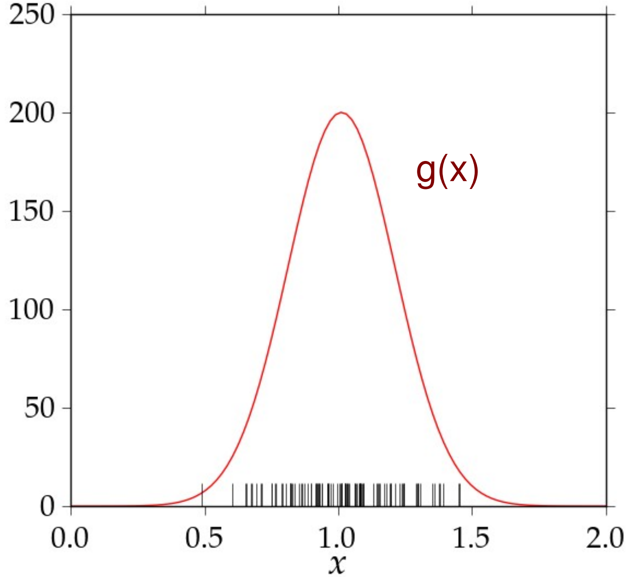
Maximizing ln*L* means solving $\partial_{\vec{p}} \ln L(\vec{p}) \overset{!}{=} 0$

Generally a non-linear problem
Minimization done numerically
(e.g. with MINUIT)

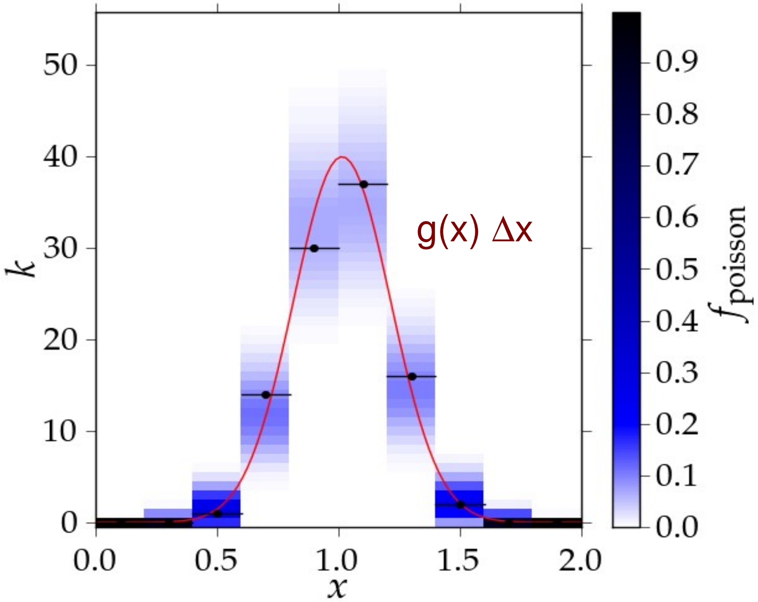
Example

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

Unbinned data



Binned data



Fit directly to point distribution

g has to be normalized

$$\ln L = \sum_i \ln[g(x_i|N,\mu,\sigma)/N] + \ln f_{\text{poisson}}(N_{\text{tot}}, N)$$

Fit to Poisson distributed histogram counts

$$\ln L = \sum_i \ln f_{\text{poisson}}(k_i, \lambda_i(N,\mu,\sigma))$$

$$\lambda_i = \int_{x_i}^{x_{i+1}} \mathrm{d}x\, f_{\text{model}}(x|N,\mu,\sigma)$$

Least-squares method

Special case of maximum-likelihood method
Only usable with binned data (or in general: x_i, y_i pairs)
Assumes **multivariate-normal distribution** of deviations from model

$$L(\vec{p}) \propto \exp\left(-\frac{1}{2}(\vec{y}-\vec{f}(\vec{x}|\vec{p}))^T \tilde{V}^{-1}(\vec{y}-\vec{f}(\vec{x}|\vec{p}))\right)$$

$x_i,\ y_i$ data pairs
$\tilde{V}_{ij} = \mathrm{cov}(y_i, y_j)$
$f_i(x_i)$ model prediction

Common case of independent observations

$$L(\vec{p}) \propto \exp\left(\sum_i \left(\frac{y_i - y(x_i|\vec{p})}{\sigma(x_i|\vec{p})}\right)^2\right)$$

Minimize $LS(\vec{p}) = -2\ln\frac{L(\vec{p})}{L(\hat{\vec{p}})} = \sum_i \left(\frac{y_i - y(x_i|\vec{p})}{\sigma(x_i|\vec{p})}\right)^2$ = sum of squared residuals
→ method of least squares

Another common simplification

Replace $\sigma(x_i|\vec{p})$ by point-wise estimates σ_i (e.g. for histogram entries $\sigma_i = \sqrt{k_i}$)

Linear least-squares method

Special case of least-squares method
Often used to get starting point for numerical minimization of LS or ML methods
Solution is unique, statistically unbiased and has minimum variance

Linear model $y(x) = \sum_j p_j \, b_j(x)$ e.g. polynomial $y(x) = p_0 + p_1 \, x + p_2 \, x^2$

$$LS(\vec{p}) = (\vec{y} - A\vec{p})^T \tilde{V}^{-1} (\vec{y} - A\vec{p}) \qquad \tilde{V}_{ij} = \mathrm{cov}(y_i, y_j) \qquad A_{ik} = b_k(x_i)$$

Minimum condition can be solved analytically

$$0 \overset{!}{=} \partial_{\vec{p}} LS = -2A^T \tilde{V}^{-1} (\vec{y} - A\,\vec{p})$$ with $$\partial_{\vec{x}}(\vec{x}^T M \vec{x}) = 2M\vec{x}, \text{ if } M^T = M$$

$$A^T \tilde{V}^{-1} \vec{y} = A^T \tilde{V}^{-1} A\,\vec{p}$$

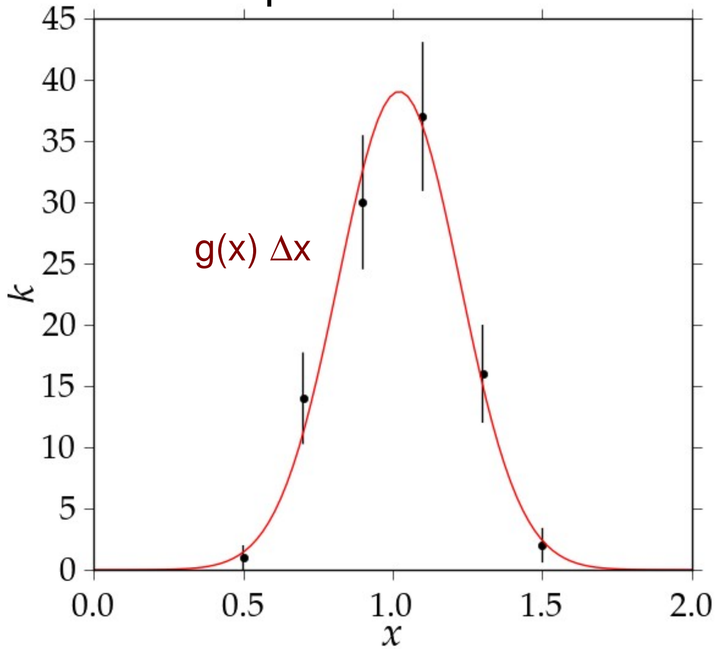
$$\vec{p} = (A^T \tilde{V}^{-1} A)^{-1} A^T \tilde{V}^{-1} \vec{y}$$

Example

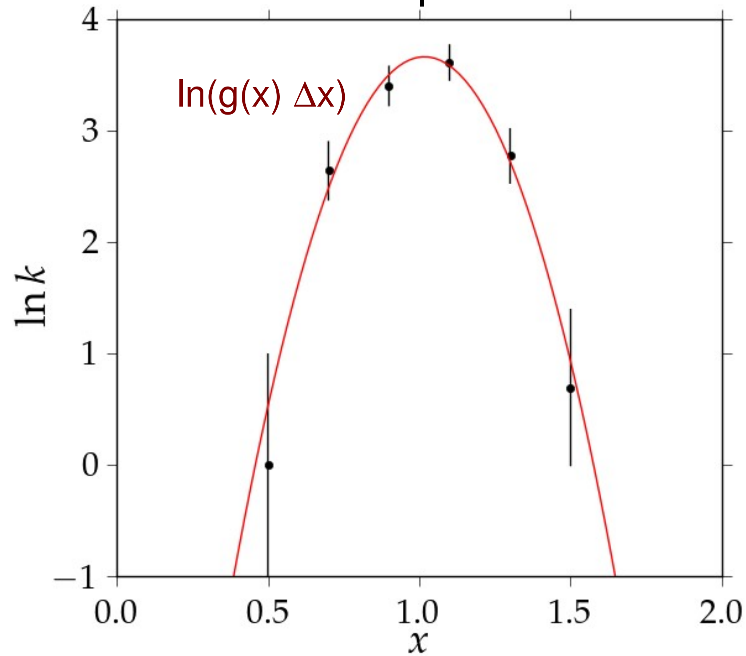
$$g(x|N,\mu,\sigma) = N\,\frac{1}{\sqrt{2\pi}\sigma}\exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$$

Fit model to histogram counts assuming normal distribution of residuals with $\sigma_i = \sqrt{y_i}$

Least-squares method



Linear least-squares method



$$LS(N,\mu,\sigma) = \sum_i \frac{(k_i - g(\bar{x}_i|N,\mu,\sigma)\Delta x)^2}{k_i}$$

Cannot use entries with $k_i = 0$
→ loss of information

$$LLS(a,b,c) = \sum_i \frac{(\ln k_i - a + b\,\bar{x}_i + c\,\bar{x}_i^2)^2}{1/\sqrt{k_i}}$$

Transform after fit $a,\ b,\ c\ \rightarrow N,\ \mu,\ \sigma$

Calculation and interpretation of fit uncertainties

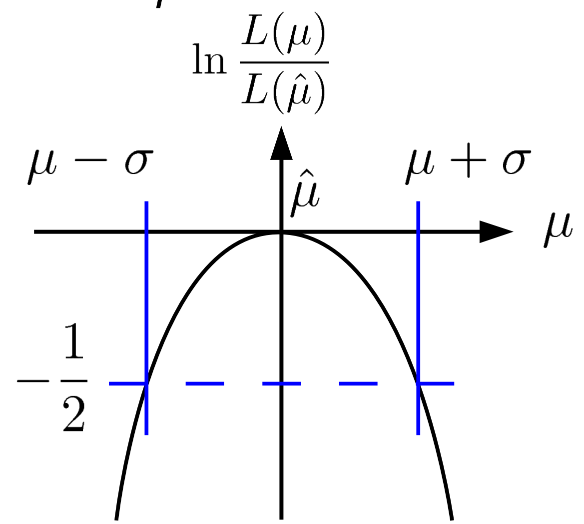
Uncertainty of ML-estimate

ln$L$ for observation $\hat{\mu}$ from normal distribution with unknown μ and known σ^2:

$$L(\mu) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(\mu-\hat{\mu})^2}{2\sigma^2}\right)$$

$$\ln\frac{L(\mu)}{L(\hat{\mu})} = -\frac{1}{2\sigma^2}(\mu-\hat{\mu})^2$$

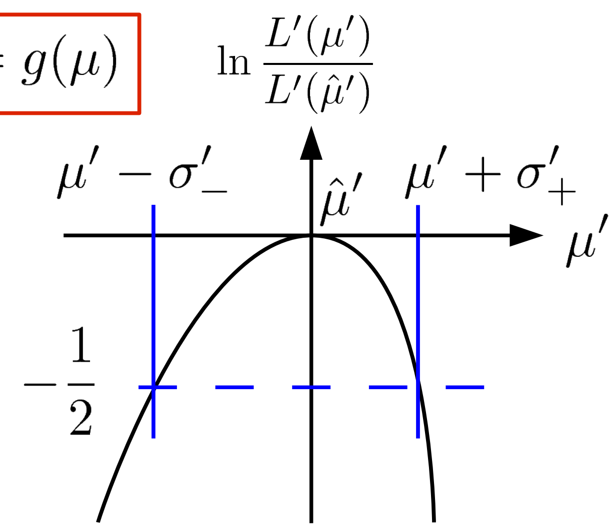
$$-\frac{1}{2} \overset{!}{=} -\frac{1}{2\sigma^2}(\mu-\hat{\mu})^2 \;\Rightarrow\; (\mu-\hat{\mu}) = \pm\sigma$$



Due to properties of normal distribution

$$P\big[-\sigma \le \mu-\hat{\mu} \le \sigma\big] = 68\,\%$$
$$P\big[\hat{\mu}-\sigma \le \mu \le \hat{\mu}+\sigma\big] = 68\,\%$$

$$\mu' = g(\mu)$$



Approach also valid in case of non-normal distribution

Invariance of likelihood ratio
$$\frac{L'(\mu')}{L'(\hat{\mu}')} = \frac{L(\mu)\,\partial\mu/\partial\mu'}{L(\hat{\mu})\,\partial\mu/\partial\mu'}$$

Uncertainty of ML-estimate

Alternative approach if $\ln L(\mu)$ is approximately parabolic

Taylor expansion around maximum

$$\ln L(\mu)\big|_{\mu=\hat{\mu}} \approx \ln L(\hat{\mu}) + \frac{1}{2}\partial_\mu^2 \ln L(\mu)\big|_{\mu=\hat{\mu}}(\mu-\hat{\mu})^2 + O(\mu-\hat{\mu})^3$$

$$\ln L(\mu) = \ln L(\hat{\mu}) - \frac{1}{2}\frac{1}{\sigma^2}(\mu-\hat{\mu})^2$$

General multivariate case

Maximum-likelihood method

$$\ln \frac{L(\vec{p})}{L(\hat{\vec{p}})} \stackrel{!}{=} -\frac{1}{2} \quad \Rightarrow \quad {p_i}^{+\sigma_i^+}_{-\sigma_i^-}$$

or

$$V \approx -\left(\partial_{p_i}\partial_{p_j} \ln L(\vec{p})\big|_{\vec{p}=\hat{\vec{p}}}\right)^{-1}$$

Least-squares method

$$LS(\vec{p}) \stackrel{!}{=} 1 \quad \Rightarrow \quad {p_i}^{+\sigma_i^+}_{-\sigma_i^-}$$

or

$$V \approx 2\left(\partial_{p_i}\partial_{p_j} LS(\vec{p})\big|_{\vec{p}=\hat{\vec{p}}}\right)^{-1}$$

Linear least-squares method

$$V = (A^T \tilde{V}^{-1} A)^{-1}$$ exact!

Bias of ML-estimate

Example: normal distribution with unknown μ, σ

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

$$\ln L(\sigma^2) \equiv -\frac{N}{2}\ln\sigma^2 - \frac{1}{2\sigma^2}\sum_i (x_i-\mu)^2 \Rightarrow 0 \stackrel{!}{=} \partial_{\sigma^2}\ln L(\sigma^2) = -\frac{N}{2\sigma^2} + \frac{1}{2\sigma^4}\sum_i (x_i-\hat{\mu})^2$$

$$\hat{\sigma}^2 = \frac{1}{N}\sum_i (x_i-\hat{\mu})^2 \quad \text{with} \quad \hat{\mu} = \frac{1}{N}\sum_i x_i$$

biased estimator of σ^2

$$E[\hat{\sigma}^2] = \frac{1}{N} N E[(x_i-\hat{\mu})^2] = E[(x_i-\mu+\mu-\hat{\mu})^2]$$

$$= E[(x_i-\mu)^2 + (\hat{\mu}-\mu)^2 - 2(\hat{\mu}-\mu)(x_i-\mu)] = \sigma^2 + \frac{\sigma^2}{N} - \frac{2\sigma^2}{N} = \sigma^2 - \frac{\sigma^2}{N}$$

$$s^2 = \frac{N}{N-1}\hat{\sigma}^2$$

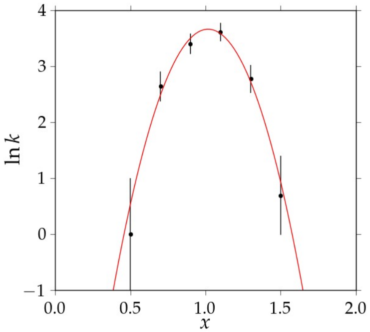
unbiased estimator of σ^2 with increased variance

$$V[s^2] = \left(\frac{N}{N-1}\right)^2 V[\hat{\sigma}^2]$$

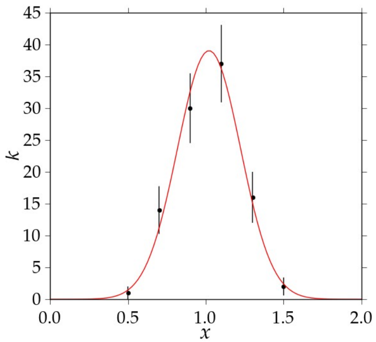
In general: ML-estimate biased if ln$L$ not parabolic $E[p-\hat{p}] \propto \partial_p^3 \ln L(p)$

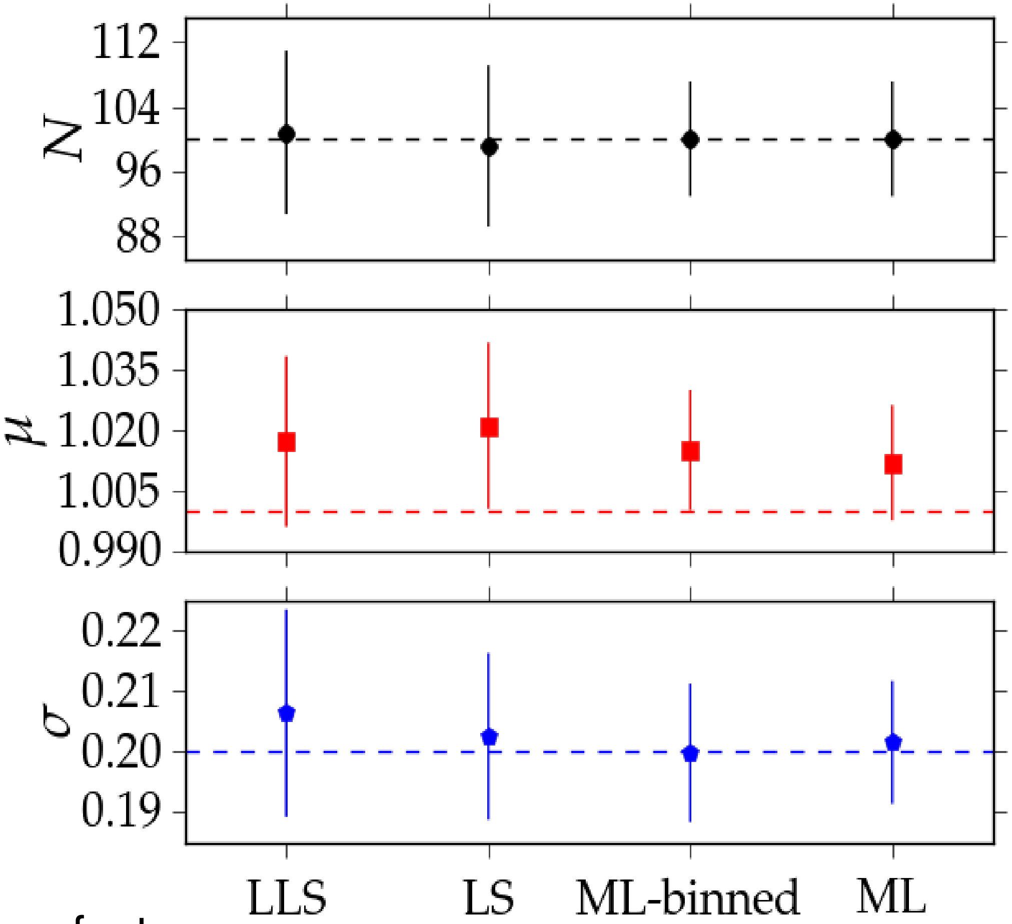
Method comparison

LLS

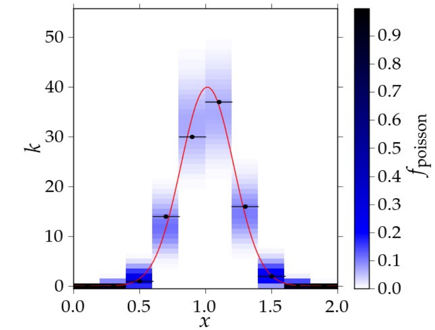


LS

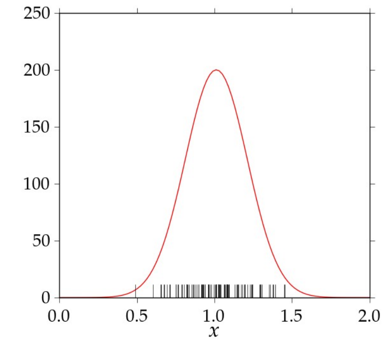




ML-binned



ML



fast
approximate
larger variance

slow
accurate
smaller variance

Coverage

How to interpret confidence regions from $\ln \frac{L(\vec{p})}{L(\hat{\vec{p}})} \stackrel{!}{=} -\frac{1}{2}$ or $LS(\vec{p}) \stackrel{!}{=} 1$?

If experiment would be repeated...

Intervals along each dimension cover true value in 68 % of all cases

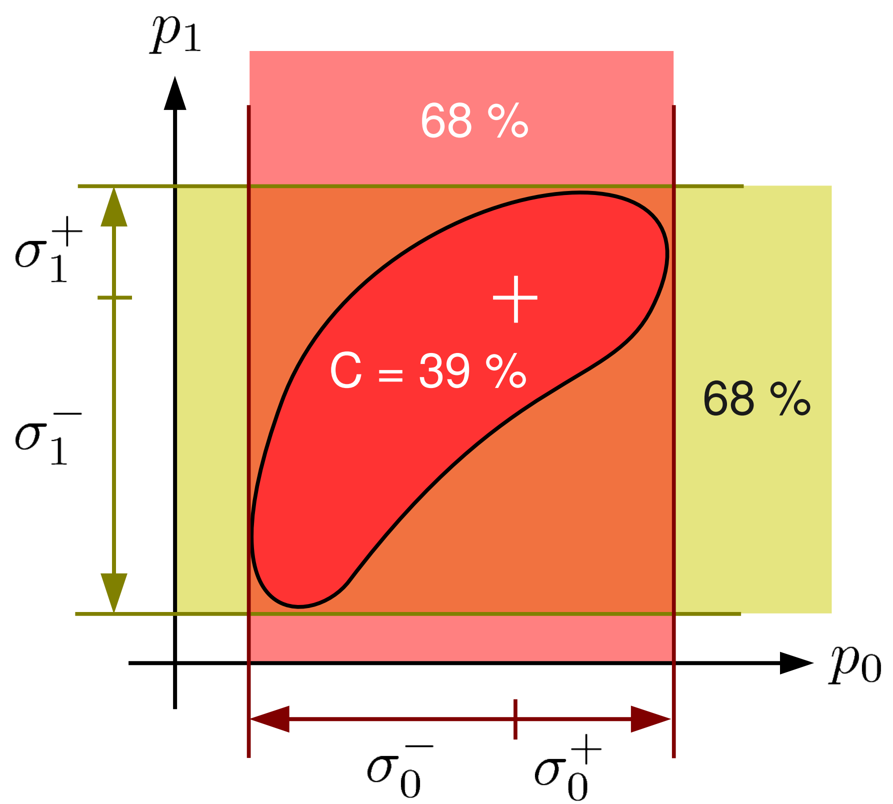
But: 2d-region covers true values only in C = 39 % of all cases

How to get C = 90 % or 99 % regions?

General case: *N* parameters
C confidence of coverage

$$\ln \frac{L(\vec{p})}{L(\hat{\vec{p}})} \stackrel{!}{=} -\frac{1}{2}\chi_\beta(C) \text{ or } LS(\vec{p}) \stackrel{!}{=} \chi_\beta(C)$$

with $\int_0^{\chi_\beta^2} \mathrm{d}x\, f_{\chi^2}(x|N) \stackrel{!}{=} C$ solved for χ_β



$N = 1,\ C = 68\,\% \rightarrow \chi_\beta \approx 1$

$N = 2,\ C = 68\,\% \rightarrow \chi_\beta \approx 1.51$

Some fitting advice

- Think carefully about the fluctuations in your problem

- Use un-binned maximum-likelihood method if possible
  - Under very general conditions, ML-estimate is asymptotically unbiased and has minimum variance (Cramer-Rao bound)

- Use linear models for empirical parametrizations
  - Fourier terms, polynomials, B-splines, …

- If you use approximate variance formula, check whether it applies

- If confidence interval is not symmetric, result is usually biased

Bayesian vs. Frequentist inference

Frequentist (Reproducability)	Bayesian (Decision theory)
Inference principle	
Likelihood function	Bayes theorem and **prior probabilities**
No treatment of systematic uncertainties	Treatment of systematic uncertainties
	"Objective Bayesian": Jeffreys or Reference priors
Point estimation	
Maximum of likelihood function	Mean of posterior probability density
Invariant to transformations	Not invariant to transformations
Interval estimation	
Based on likelihood ratio	Quantiles of posterior probability density
Coverage	Credible interval tells nothing about coverage
Restriction of a parameter at a physical boundary	
Via parameter transformation	Via prior probabilities