Some More Statistical Concepts & Terms Relevant in Gamma-Ray Astronomy*

Ullrich Schwanke
Humboldt-Universität zu Berlin

*** Not only, of course ☺**

Motivation

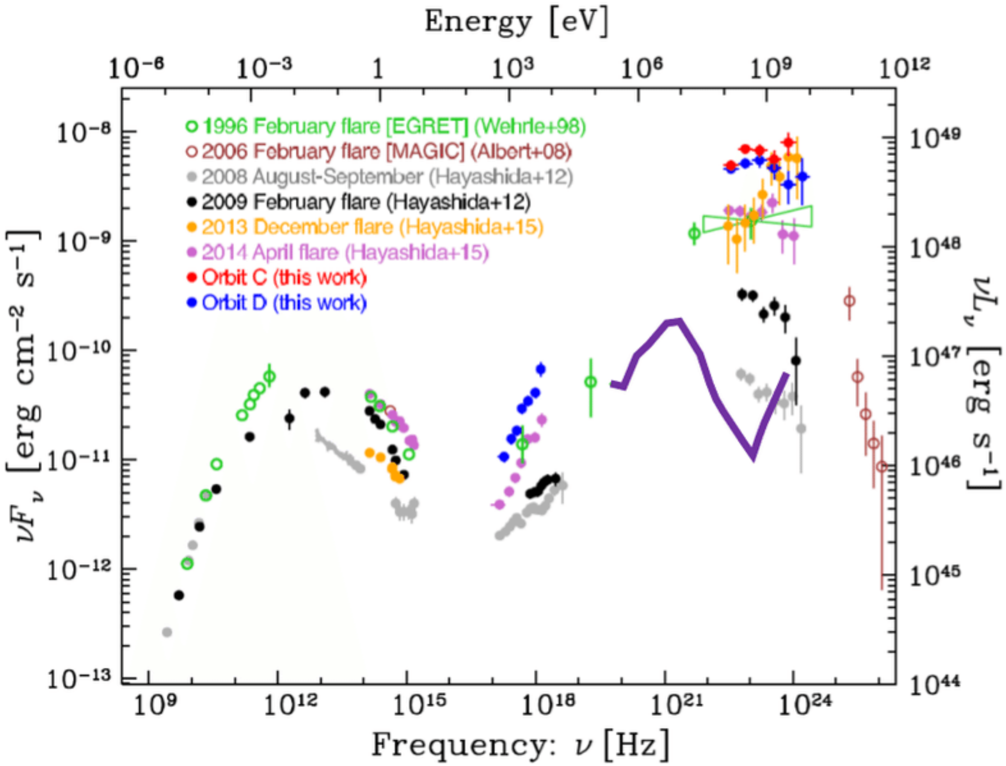
- **Glen's lectures have introduced unbinned maximum-likelihood estimators and their uses in**
  - **Point estimation**
  - **Interval estimation**
  - **Hypothesis testing**
- **In the end, the likelihood of an experiment encodes all the information – but it is not always accessible for outside people**
- **Would like to elaborate on a few concepts and terms that are also relevant for gamma-ray astronomy (and the workshop starting tomorrow)**

Content

- **Error propagation/change of variables**
- **Statistical and systematic errors**
- **Binned maximum likelihood and model testing**
- **Trial factors /look-elsewhere effect**

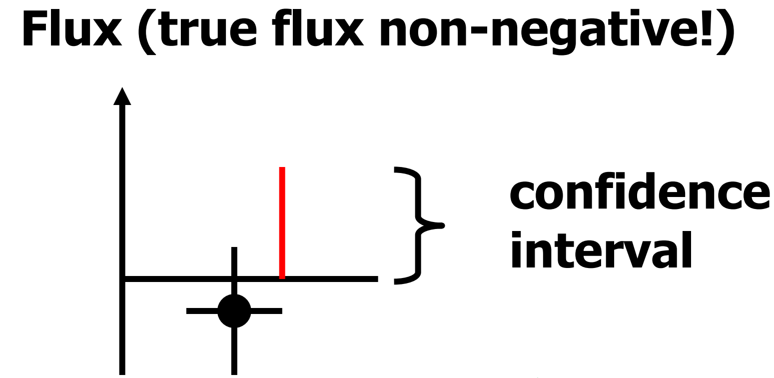
Content

- **Error propagation/change of variables**
- **Statistical and systematic errors**
- **Binned maximum likelihood and model testing**
- **Trial factors /look-elsewhere effect**



Variance and Confidence Intervals

Measurements within ±1σ around mean	
Gauss	68.3%
Exponential	86.5%
Uniform distribution	57.7%

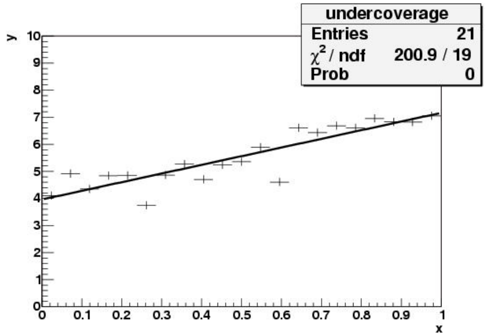


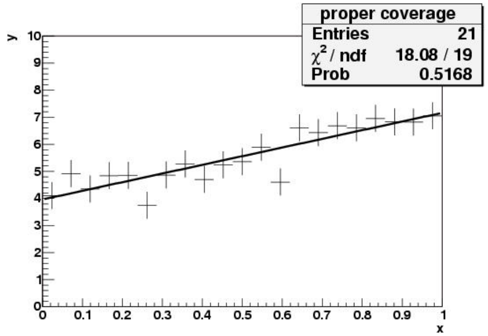
$$V(x_1) = E\left[(x_1 - E[x_1])^2\right]$$

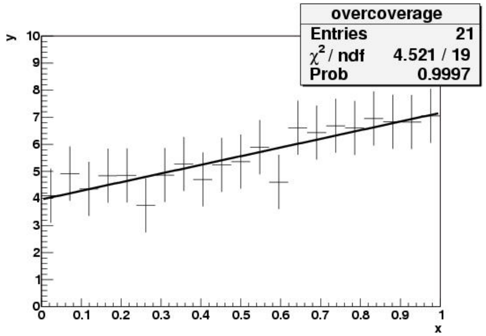
random variable PDF $f(x_1|\theta)$

- **Lecture 1: sqrt(Variance) as measure of the width of a PDF**
- **This „error“ is not accurate enough („Probability content“ depends on shape/type of PDF)**
- **A confidence interval (CI) should (i) include the true value of a parameter with some probability (degree of belief, Bayesian) or (ii) belong to an ensemble of CIs a certain fraction of which (confidence level) includes the true value (Frequentist)**

Confidence Intervals: Coverage







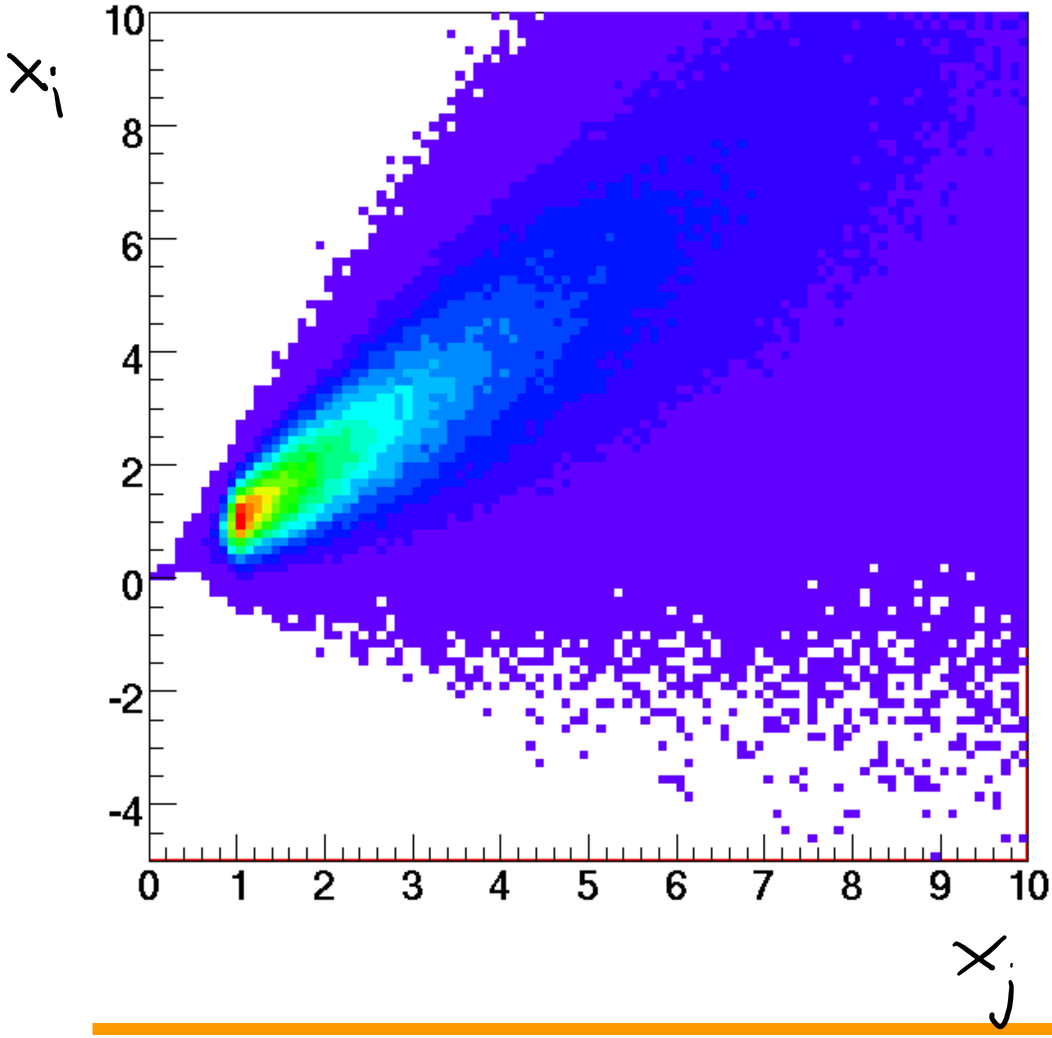
- Confidence intervals too narrow "undercoverage"
- Measurement appears more precise than it is (should be avoided)

- Correct coverage

- Confidence intervals too broad (i.e. too conservative) "overcoverage"
- Excludes fewer (wrong) hypotheses

- Proper coverage of calculated confidence intervals can be tested with the help of Monte Carlo simulations (see appendix for pseudo codes for the Frequentist and Bayesian case)

Variance and Covariance



- **PDF $f(x_1,x_2|\theta)$**

$$E[x_i] = \langle x_i \rangle = \bar{x}_i = \int dx_1 \int dx_2 \, x_i f(x_1,x_2|\theta)$$

Variance and Covariance

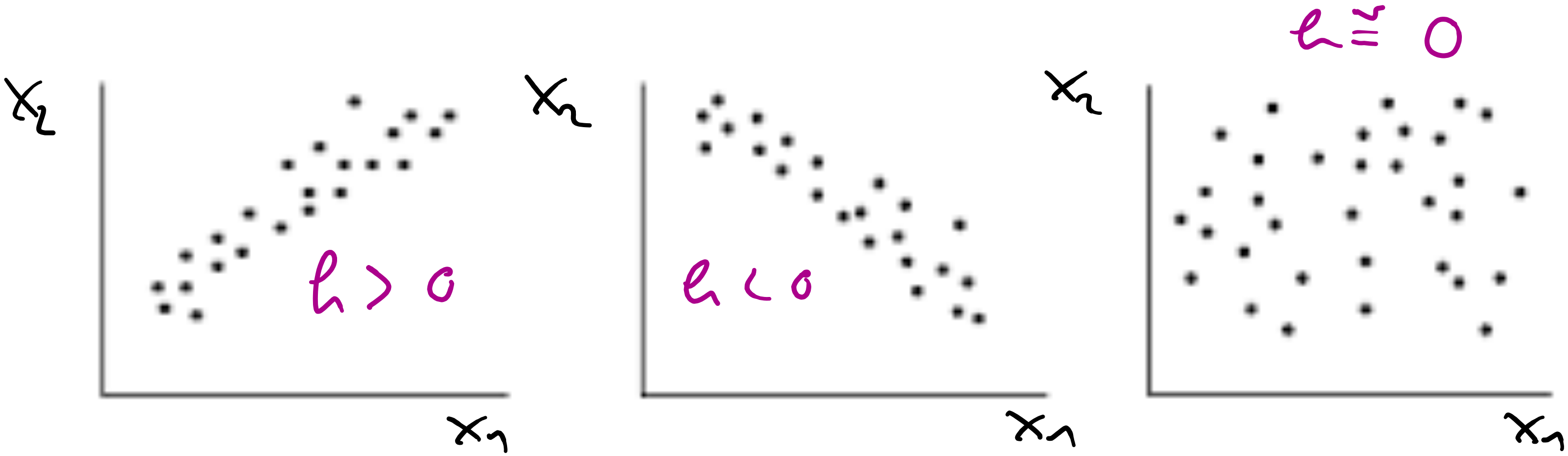
PDF $f(x_1, x_2 | \vec{\theta})$

$$\mathrm{cov}(x_i x_j) = E[(x_i - \bar{x}_i)(x_j - \bar{x}_j)]$$

$$V(x_1) = E[(x_1 - E[x_1])^2] = \mathrm{cov}(x_1, x_1)$$

- **If PDF $f(x_1,x_2)$ factorizes as $f(x_1,x_2) = f_1(x_1)\, f_2(x_2)$ the random variables are mutually independent and their covariance is 0**
- **Important: The converse statement is not true (i.e. one cannot claim that two variables are independent if their covariance vanishes)**
- **For N variables, $\mathrm{cov}(x_1,..,x_N)$ is a symmetric NxN matrix that is called covariance matrix/variance matrix/error matrix**

Correlation Coefficient



$$\rho_{ij} = \frac{\mathrm{cov}(x_i, x_j)}{\sqrt{\mathrm{cov}(x_i, x_i)\,\mathrm{cov}(x_j, x_j)}}$$

$$V(\vec{x}) = \begin{pmatrix} \sigma_1^2 & \rho_{12}\sigma_1\sigma_2 \\ \rho_{12}\sigma_1\sigma_2 & \sigma_2^2 \end{pmatrix}$$

- $\rho_{ii} = 1$
- $-1 \leq \rho \leq +1$

Covariance Matrix (1/2)

$$V[\hat{\theta}] \geq \left(1 + \frac{\partial b}{\partial \theta}\right)^2 \Big/ E\left[-\frac{\partial^2 \ln L}{\partial \theta^2}\right]$$ = MVB (Minimum Variance Bound)

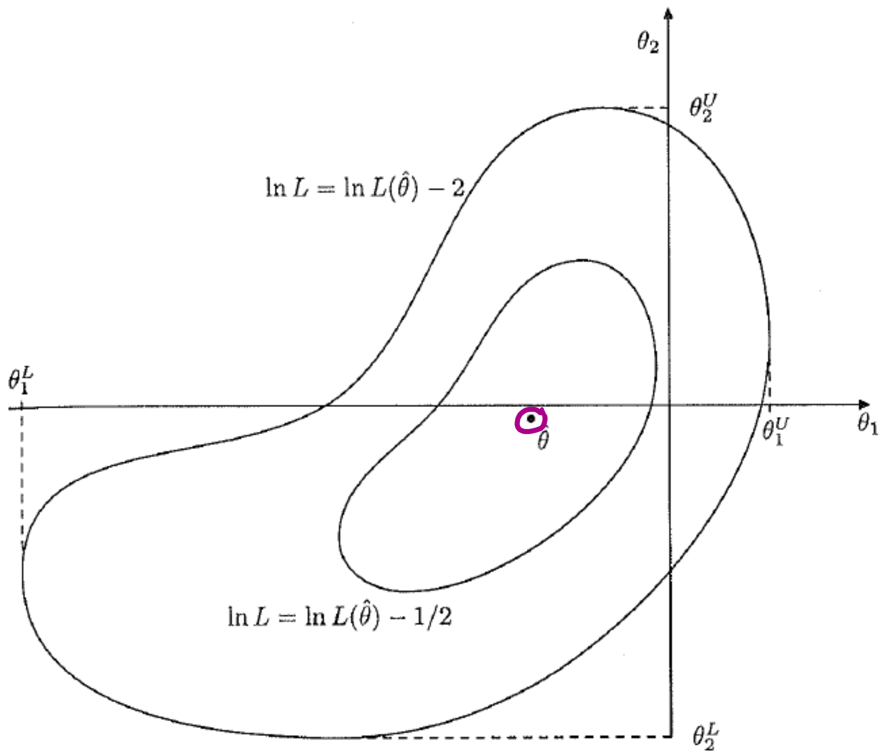
e.g. 2 parameters $i,j = 1 \ldots 2$

$$V(\hat{\theta}_{ij})^{-1} = -E\left[\frac{\partial^2 \ln L}{\partial \theta_i \partial \theta_j}\right]$$

$$= -\frac{\partial^2 \ln L}{\partial \theta_i \partial \theta_j}\Bigg|_{\theta_i = \hat{\theta}_i,\ \theta_j = \hat{\theta}_j}$$

- **In the ML scheme, the covariance matrix can be estimated (often numerically) from the Hessian Matrix of 2nd derivatives**
- **Strictly valid (only) in the limit of large N**
- **PDF of the estimate is then a multivariate Gaussian, no bias**

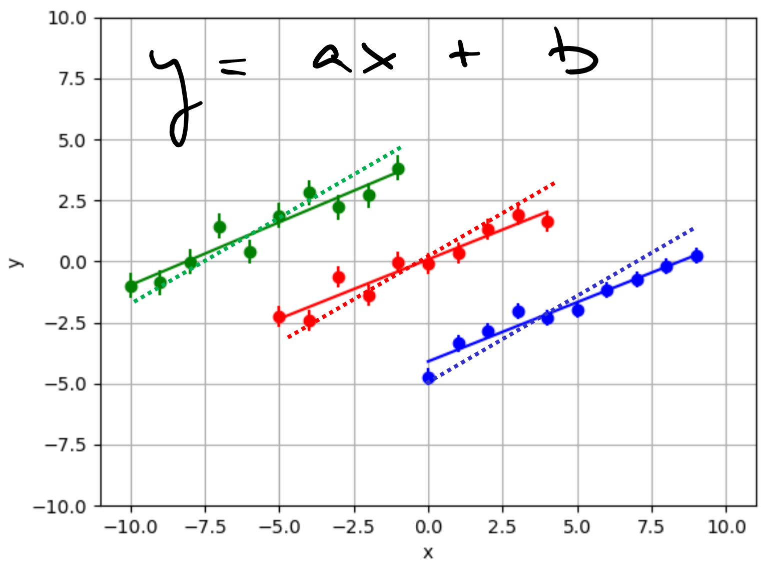
Covariance Matrix (2/2)



Contours at confidence levels of 39.9% and 86.5%

- **The estimate of the covariance matrix from the derivatives near the optimal parameter values is an approximation**
- **Approximation will be bad when the likelihood is still non-Gaussian**
- **The likelihood encodes more information than the covariance matrix (unless N→∞)**

Removing Correlation (1/2)



- cov(a,b)>0
- cov(a,b)=0
- cov(a,b)<0

$$L = \prod_{k=1}^{N} \frac{e^{-\frac{1}{2}\left(\frac{y_k-(ax_k+b)}{\sigma_k}\right)^2}}{\sqrt{2\pi}\,\sigma_k}$$

- **Cov(i,j)-terms (i≠j) can be brought to zero by a suitable transformation**
- **The transformation will introduce new parameters that one has to cite in connection with revised covariance matrix**
- **Common application: decorrelation energy when fitting spectral models (flux as a function of energy)**

Removing Correlation (2/2)

$$-\ln L = \frac{1}{2}\sum_k \left[\frac{y_k - (a x_k + b)}{\sigma_k}\right]^2 + \text{const.}$$

$$-\frac{\partial \ln L}{\partial a} = \sum_k \left[\frac{y_k - (a x_k + b)}{\sigma_k}\right]\left(-\frac{x_k}{\sigma_k}\right)$$

$$-\frac{\partial^2 \ln L}{\partial b\, \partial a} = \sum_k \left(-\frac{1}{\sigma_k}\right)\left(-\frac{x_k}{\sigma_k}\right) = \sum_k \frac{x_k}{\sigma_k^2}$$

transformation:

$$\frac{x_k'}{\sigma_k^2} = \frac{x_k}{\sigma_k^2} - \frac{1}{N}\sum \frac{x_k}{\sigma_k^2}$$

- **ML and Least Squares are equivalent when the PDF is Gaussian**
- **Recall: Inverting a symmetric 2D matrix scales the off-diagonal element and changes its sign**

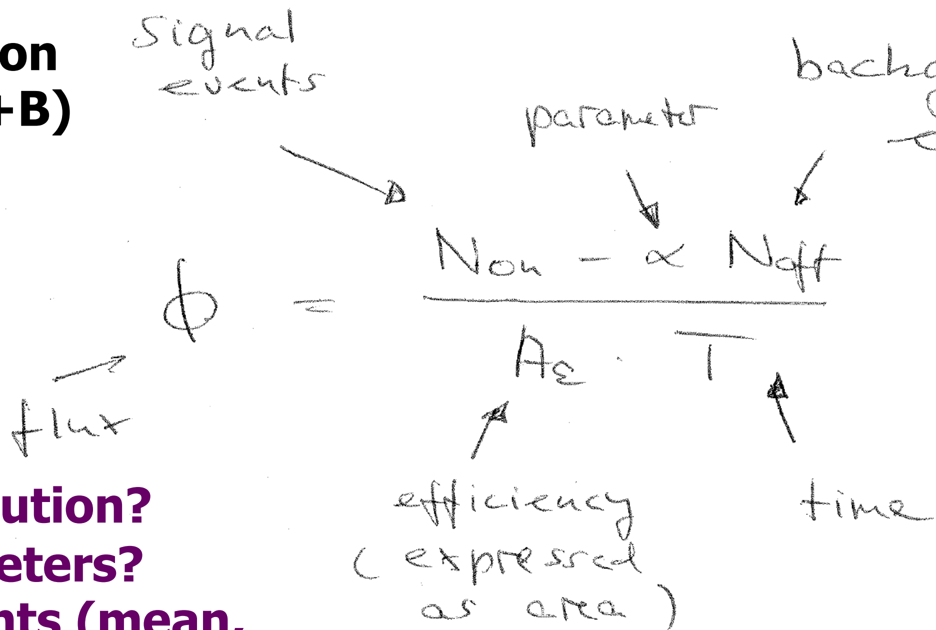
Change of Variables

Transformation (RV=random variable)	New PDF
Sum of N x_i^2 when x_i is RV from Norm(0,1)	$\chi^2(N)$
Sum of N $\alpha_i x_i$ when x_i from Norm(0,1), α_i constant	Gaussian
Quotient of x_1 and x_2, both from Norm(0,1)	Cauchy distribution
Sum of two RV from U(0,1)	Triangular distribution

- **Suppose x follows PDF $f_x(x|\theta)$ and we apply y = f(x)**
- **Would like to know the PDF $f_y(y|\theta`)$ for y and the mapping from $\theta$ to $\theta`$**
- **Old and new PDF are known for some cases (→table), and the concept of a characteristic function and transformation formulae are helpful when deriving the new PDF**
- **In (experimental) practice, the x is a vector of variables and the PDF will anyway be quite complicated when one folds in effects like (energy, space) resolution and acceptance**

Error Propagation

Poisson distribution (mean S+B)



Poisson distribution (mean B/α)

Distribution?
Parameters?
Moments (mean, variance) ?

Uniform (least significant bit..)

Binomial (n out of N simulated events)

- **Error propagation is a term used by experimentalists**
- **Error propagation is approximate change of variables**

Approximate Error Propagation

$$y = f(x_1 \ldots x_N) \qquad E[x_i] = \mu_i$$

$$\cong \underbrace{f(\mu_1 \ldots \mu_N)}_{E[y]} + \sum_{i=1}^{N} \left.\frac{df}{dx_i}\right|_{\mu_i} (x_i - \mu_i) \ldots$$

scalar

$$V(y) = E[(y - E[y])^2] = \sum_{i,j=1}^{N} \left.\frac{df}{dx_i}\right|_{\mu_i} \left.\frac{df}{dx_j}\right|_{\mu_j} E[(x_i-\mu_i)(x_j-\mu_j)]$$

$$V(y) = B\,V(x)\,B^T, \qquad B = \left.\left(\frac{df}{dx_1}, \ldots \frac{df}{dx_N}\right)\right|_{\vec{\mu}}$$

variance

$N \times N$ covariance matrix

Approximate Error Propagation

$$y_l = f_l(x_1 \dots x_N) \qquad E[x_i] = \mu_i$$

$$\cong \underbrace{f_l(\mu_1 \dots \mu_N)}_{E[y]} + \sum_{i=1}^{N} \left.\frac{df_l}{dx_i}\right|_{\mu_i} (x_i - \mu_i) \dots$$

vector $y_1 \dots y_M$

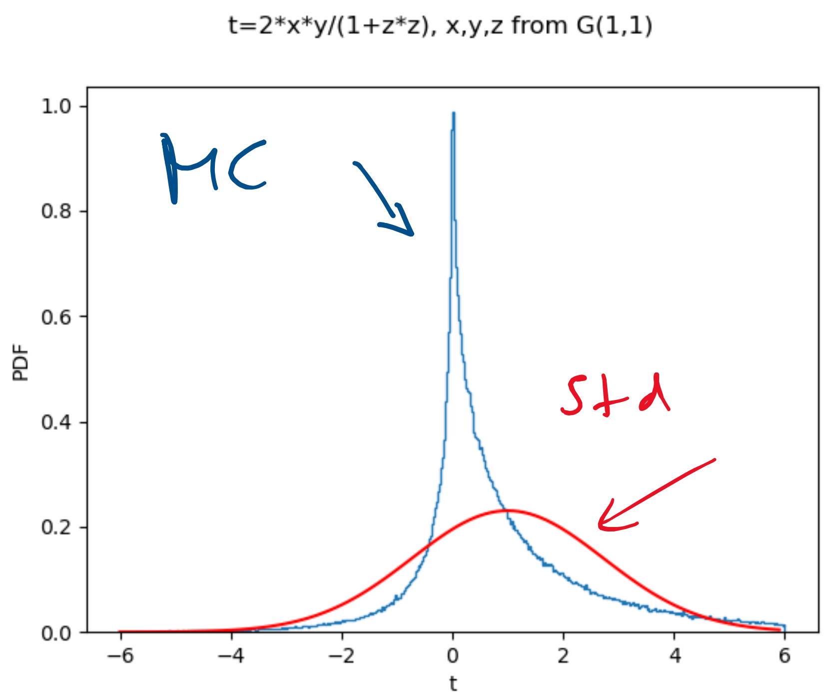
$$V(y) = E[(y_l - E[y_l])(y_k - E[y_k])] = \sum_{i,j=1}^{N} \frac{df_l}{dx_i}\frac{df_k}{dx_j} E[(x_i-\mu_i)(x_j-\mu_j)]$$

$$V(y) = B\, V(x)\, B^T$$

$M \times M$ $\qquad$ $N \times N$

$$B = \begin{pmatrix} \frac{\partial f_1}{\partial x_1} & \dots & \frac{\partial f_1}{x_N} \\ \vdots & & \\ \frac{\partial f_M}{\partial x_1} & \dots & \frac{\partial f_M}{\partial x_N} \end{pmatrix}$$

MC Error Propagation



Example: $t = \frac{2xy}{1+z^2}$

x,y,z from $G(1,1)$

$\left.\frac{dt}{dx}\right|_1 = \left.\frac{dt}{dy}\right|_1 = \left.\frac{dt}{dz}\right|_1 = 1$

$\hookrightarrow \sigma^2 = 1^2 + 1^2 + 1^2$

- **Standard error propagation is only approximate except in the linear case**
- **Sampling the input distribution with MC techniques is often an alternative**

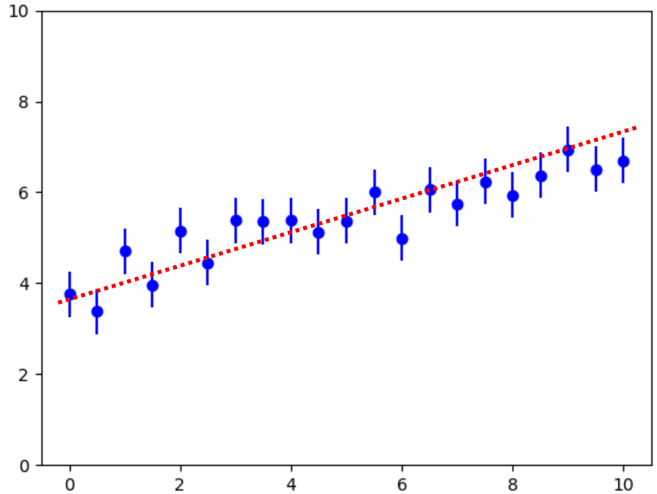
Example Credit: Michael Schmelling

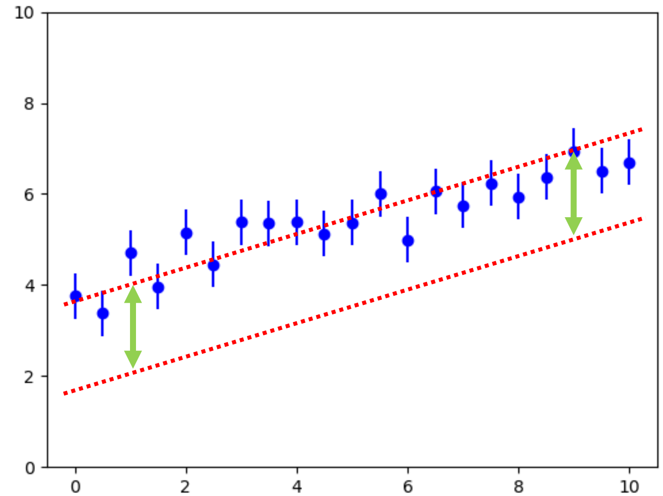
Content

- **Error propagation/change of variables**
- **Statistical and systematic errors**
- **Binned maximum likelihood and model testing**
- **Trial factors /look-elsewhere effect**

Errors: Statistical vs Systematic

True values





True values

- **Statistical errors:**
- **Deviations to lower and higher values**
- **Precision improves with 1/sqrt(N)**
- **PDF (mostly) known**

- **Systematic errors:**
- **Deviations into the same direction**
- **Repeated measurements (at the same point) are not independent; central limit theorem does not apply**
- **In fact, all measurements are correlated**
- **PDF (mostly) unknown**

Systematic Errors

- **Statistical errors can become systematic ones**
- **Systematic errors can become statistical ones (randomizing the sequence of data)**
- **There are obvious techniques to avoid systematic errors (e.g. to measure ratios)**
- **Can be identified with suitable methods (conservation laws, measure a quantity as a function of a variable it should not depend on)**
- **Systematic errors and statistical error occur independently**
- **Systematic errors can be treated with the usual statistical methods**

$$\phi = \frac{N_{on} - \alpha N_{off}}{A_\varepsilon \cdot T}$$

flux; efficiency (expressed as area); time

$$z_1 = x_1 + y_1$$
$$z_2 = x_2 + y_2$$

stat (x_1, x_2): independent, σ_1^2, σ_2^2

syst (y_1, y_2): 100% correlated, s^2

$$\Rightarrow \mathrm{cov}(x_i, y_i) = 0$$

$$i = 1, 2$$

Example (1/2)

$$C^{stat} = \begin{pmatrix} \sigma_1^2 & 0 \\ 0 & \sigma_2^2 \end{pmatrix} \qquad z_1 = x_1 + y_1, \quad z_2 = x_2 + y_2$$

$$V(z_1) = E\left[\{(x_1 - \bar{x}_1) + (y_1 - \bar{y}_1)\}^2\right]$$

$$= \sigma_1^2 + s^2 + \underbrace{2\,\mathrm{cov}(x_1, y_1)}_{0}$$

$$\mathrm{cov}(z_1, z_2) = E\left[\left((x_1 - \bar{x}_1) + (y_1 - \bar{y}_1)\right)\left((x_2 - \bar{x}_2) + (y_1 - \bar{y}_1)\right)\right]$$

$$= E\left[(y_1 - \bar{y}_1)(y_2 - \bar{y}_2)\right] = s^2$$

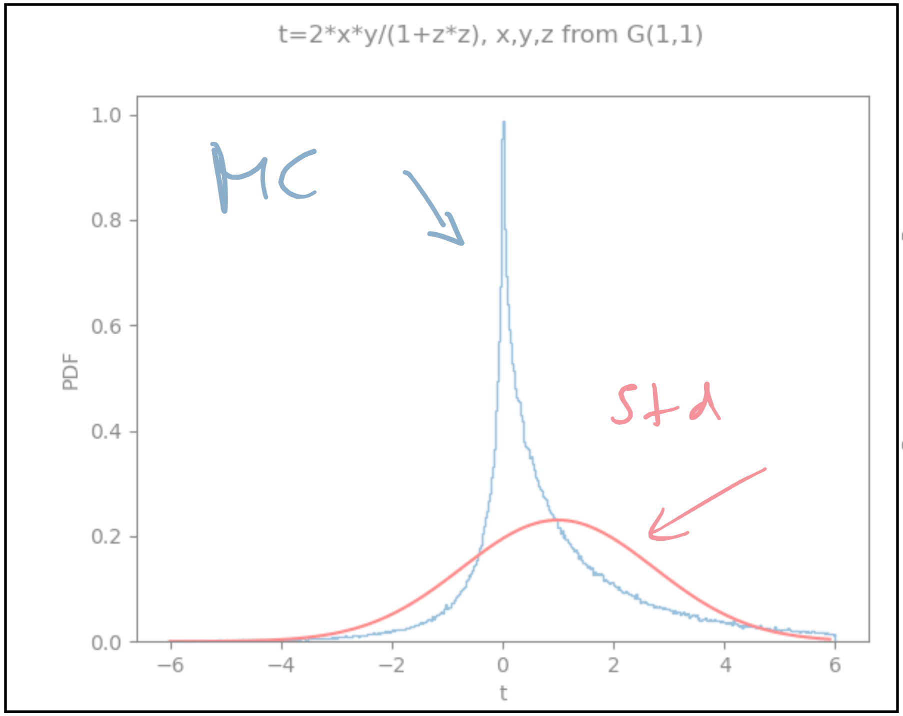
$$C^{syst} = \begin{pmatrix} s^2 & s^2 \\ s^2 & s^2 \end{pmatrix}$$

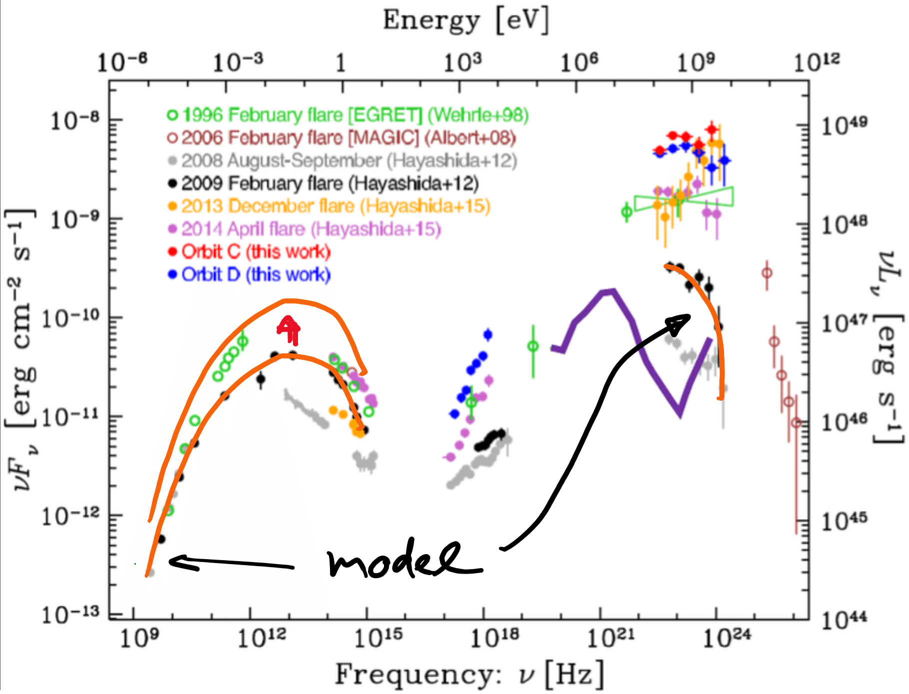
$$C = C^{stat} + C^{syst} = \begin{pmatrix} \sigma_1^2 + s^2 & s^2 \\ s^2 & \sigma_2^2 + s^2 \end{pmatrix}$$

Example (2/2)

$$V(z_1 \pm z_2) = \sigma_1^2 + \sigma_2^2 + 2(s^2 \pm s^2)$$

MC Error Propagation: Systematics

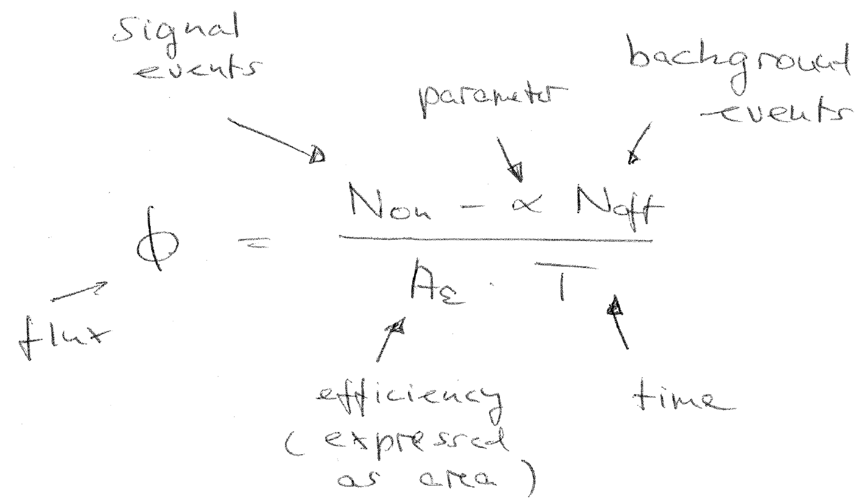




- Systematic errors are likely correlated for the same experiment/observatory but not between different experiments/observatories
- Vary all points of an experiment in the same direction....

Including Systematics

- **The presence of systematic errors („nuisance parameters") must broaden confidence intervals**
- **There are a number of Bayesian/Frequentist/hybrid procedures the (Frequentist) coverage of which is tested with the help of simulations**
- **Maximum Likelihood errors with the profile likelihood method have become a standard**



$$P(n|s,b) = \frac{(s+b)^n}{n!} e^{-(s+b)}$$

$$\tilde{P}(n|s,b) \sim \int_0^\infty P(n|s,b')\, e^{-\frac{1}{2}\left(\frac{b-b'}{\sigma_b}\right)^2} db'$$

syst. error of b

Profile Likelihood

- **CI for single parameters of interest (e.g. π) can be obtained by constructing a likelihood ratio that depends only on this parameter (1 degree of freedom)**
- **All others parameters $\theta_1, \ldots, \theta_k$ are maximised at all times for the given value $\pi = \pi_0$**
- **Of course, this also works for a parameter space $\pi_{1,\ldots,}\pi_n$**

Parameters of interest (mass, flux)

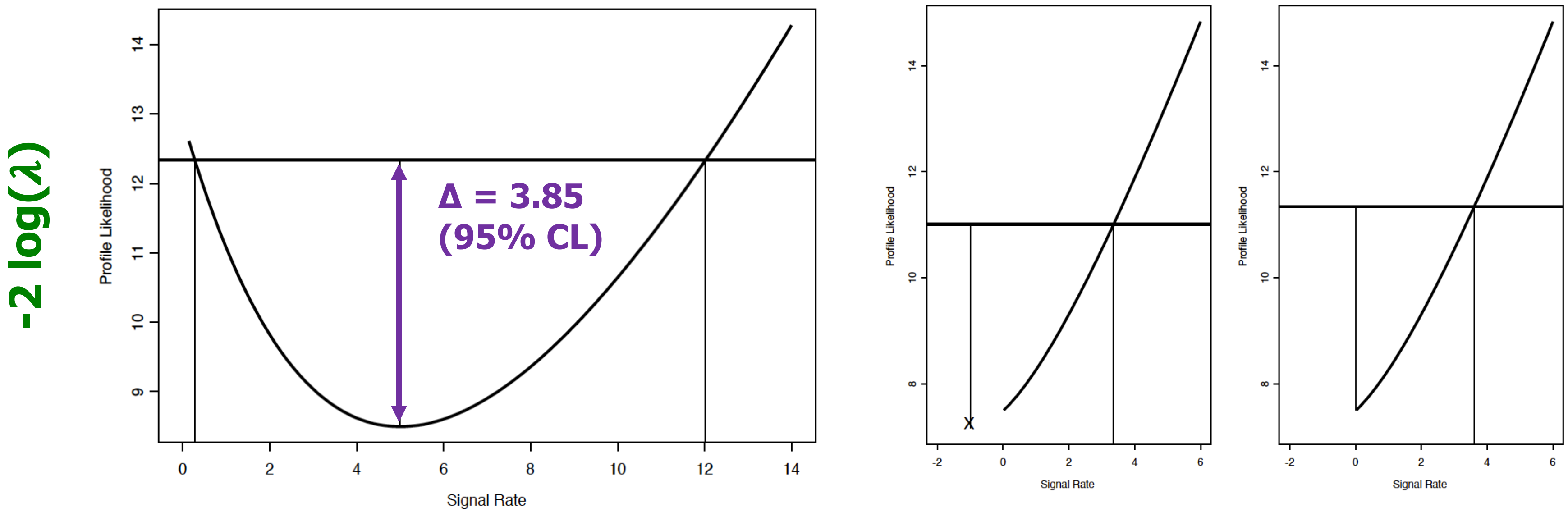
$$L(\boldsymbol{\pi}, \boldsymbol{\theta}|X) = \prod_{i=1}^{n} f(X_i|\boldsymbol{\pi}, \boldsymbol{\theta})$$

nuisance parameters (efficiency, constants)

Profile likelihood:

$$\lambda(\boldsymbol{\pi}_0|X) = \frac{\sup\{L(\boldsymbol{\pi}_0, \boldsymbol{\theta}|X); \boldsymbol{\theta}\}}{\sup\{L(\boldsymbol{\pi}, \boldsymbol{\theta}|X); \boldsymbol{\pi}, \boldsymbol{\theta}\}}$$

Profile Likelihood

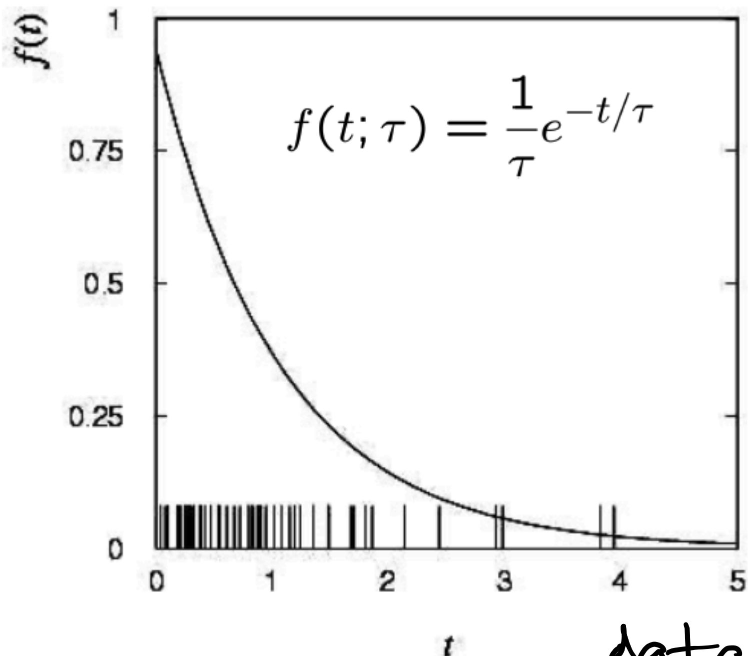


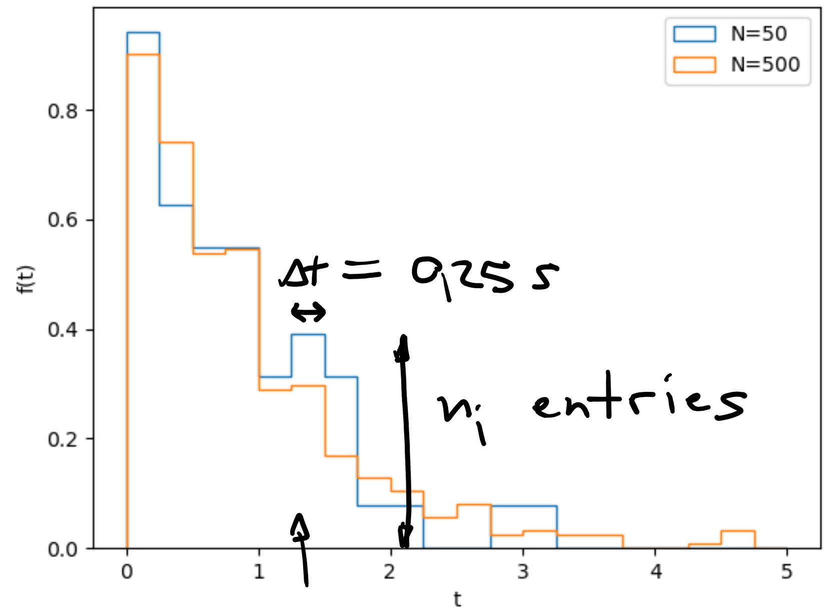
- Ad-hoc prescriptions when (i) the minimum is in the unphysical range or (ii) when the required increase leads into the unphysical range
- Important for small N, when log(L) can be highly non-Gaussian
- Software tuned to give proper coverage; provides several PDFs for data and efficiency etc (see e.g. arXiv:0403059)

Content

- **Error propagation/change of variables**
- **Statistical and systematic errors**
- **Binned maximum likelihood and model testing**
- **Trial factors /look-elsewhere effect**

ML: Unbinned vs Binned



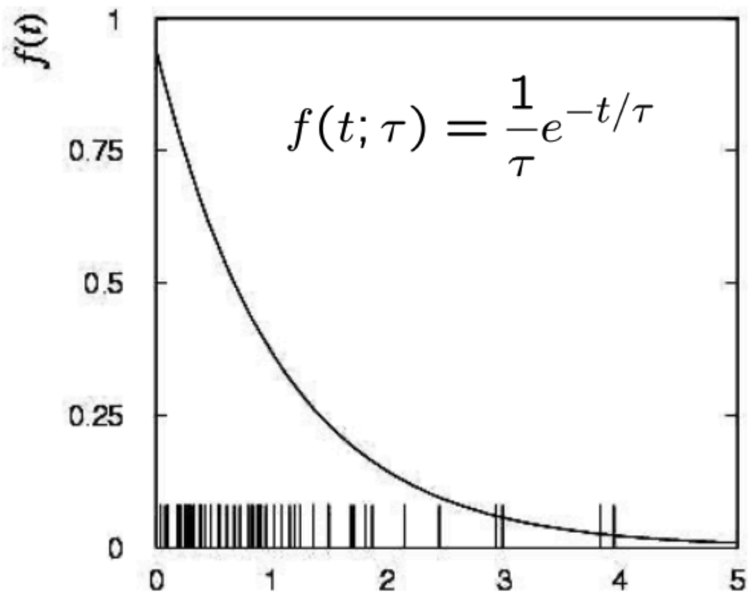


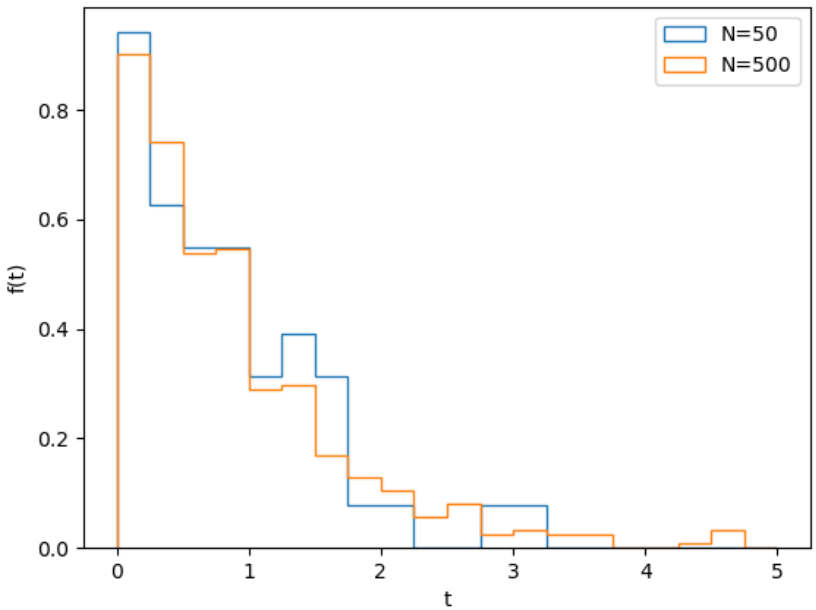
data — model

$$f_i(n_i|\tau) \sim \frac{\lambda_i^{n_i}}{n_i!} e^{-\lambda_i} \qquad \lambda_i \sim \int_{\text{bin } i} e^{-t/\tau}\, dt$$

λ_i **average number of counts expected in bin i**

ML: Unbinned vs Binned

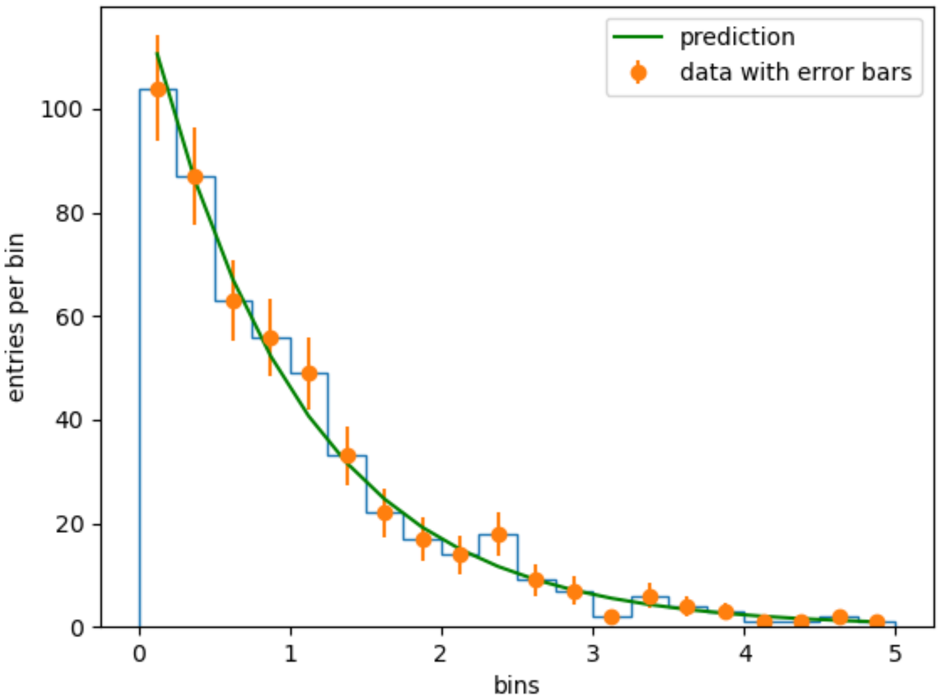




- No loss of information due to binning effects
- Number of terms in L ~ events
- Goodness of fit testing might require a binning anyway

- Loss of information due to binning effects (horrible in this example!)
- Number of terms in L ~ bins
- Goodness of fit testing basically straightforward

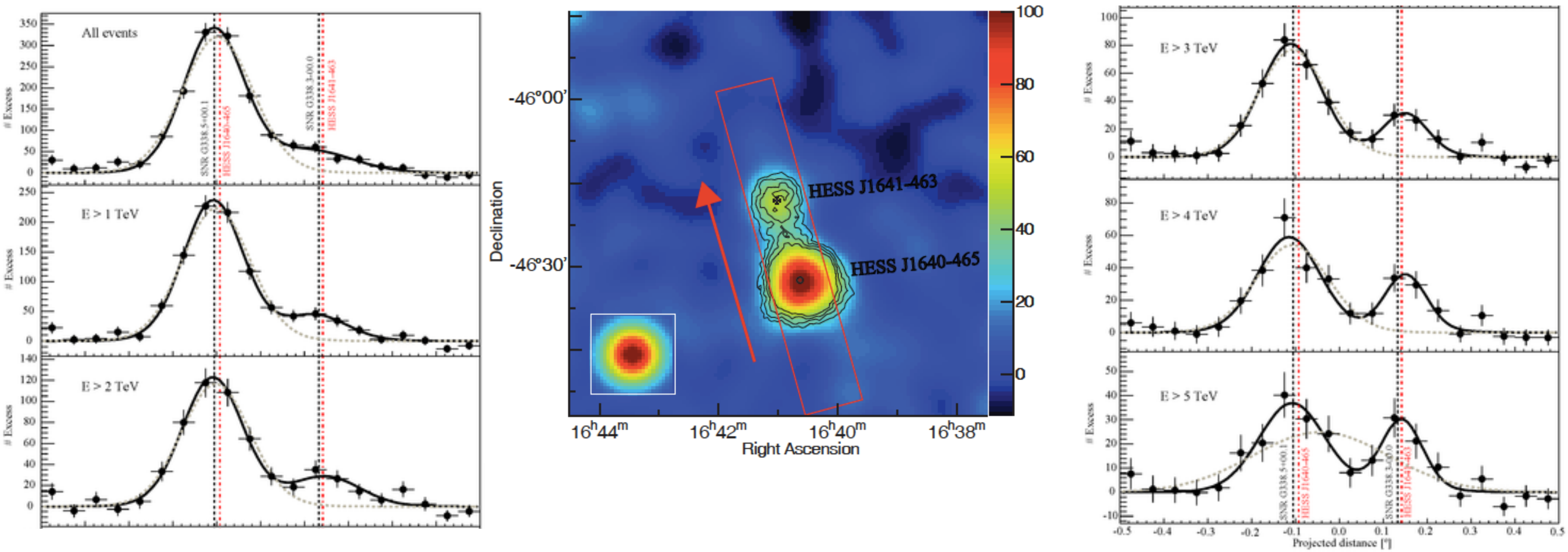
Goodness of Fit



$$\chi^2 = \sum_{i=1}^{bins} \left(\frac{n_i - n_i^{pred}}{\sigma_i} \right)^2$$

- Number of degrees of freedom (Ndof) is number of histogram bins if the prediction (=model) is completey defined
- Ndof is decreased by M if M model parameters are estimated from the binned data
- Ndof unclear when the M model parameters are estimated by unbinned ML
- Note: Assume that the bin content is “high enough“ for Gaussian approximation

Binned Maximum Likelihood



- **Binned ML is popular due to a high number of bins (spatial bins, energy bins), the desire for automatization (e.g. catalogue production and data modelling) and the intimate relation with GOF**
- **The value of -2 log(likelihood) at the maximum is asymptotically Chi^2 distributed and can be used in tests immediately**

Binned Maximum Likelihood

$$L = \prod_{i=1}^{N} e^{-x_i} \frac{x_i^{n_i}}{n_i!}$$

n_i : data

x_i : model

$$\hookrightarrow \quad -2 \log L = 2 \sum_i x_i - n_i \log x_i + \underbrace{\log n_i!}$$

Stirling (n_i large): $n_i \log n_i - n_i$

$$= 2 \sum_i (x_i - n_i) - n_i \underbrace{\log (x_i / n_i)}$$

$$x \cong 0$$

$$\log \left(1 + \underbrace{\boxed{\frac{x_i - n_i}{n_i}}}\right) \cong x - \frac{1}{2} x^2$$

$$-2 \log L = 2 \sum (x_i - n_i) - (x_i - n_i) + \frac{1}{2} \frac{(x_i - n_i)^2}{n_i}$$

Binned Maximum Likelihood

$$-2\log L \simeq \sum_i \left(\frac{x_i - n_i}{\sqrt{n_i}}\right)^2$$

$$\triangleright \quad \frac{1}{n_i} = \frac{1}{x_i + (n_i - x_i)} \simeq \left(\frac{1}{x_i} - \frac{(n_i - x_i)}{x_i^2}\right)$$

small

$$= \frac{1}{x_i}\left(1 - \frac{n_i}{x_i} + 1\right)$$

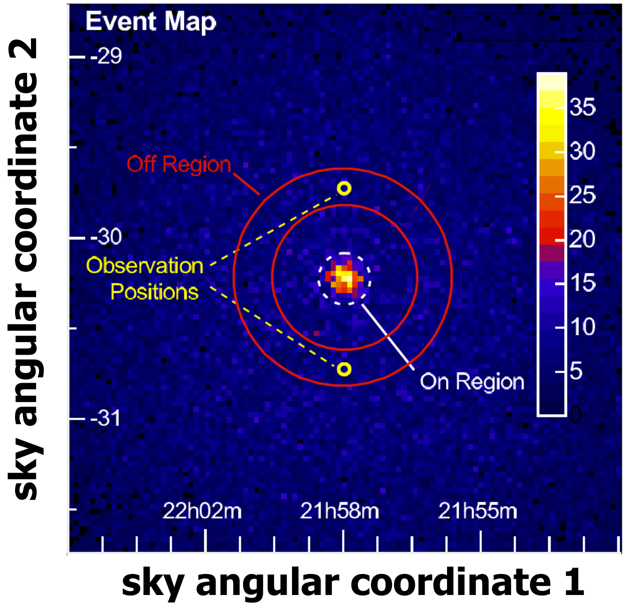
$$-2\log L \simeq \sum_i \left(\frac{x_i - n_i}{\sqrt{x_i}}\right)^2 \left(1 - \frac{n_i}{x_i} + 1\right)$$

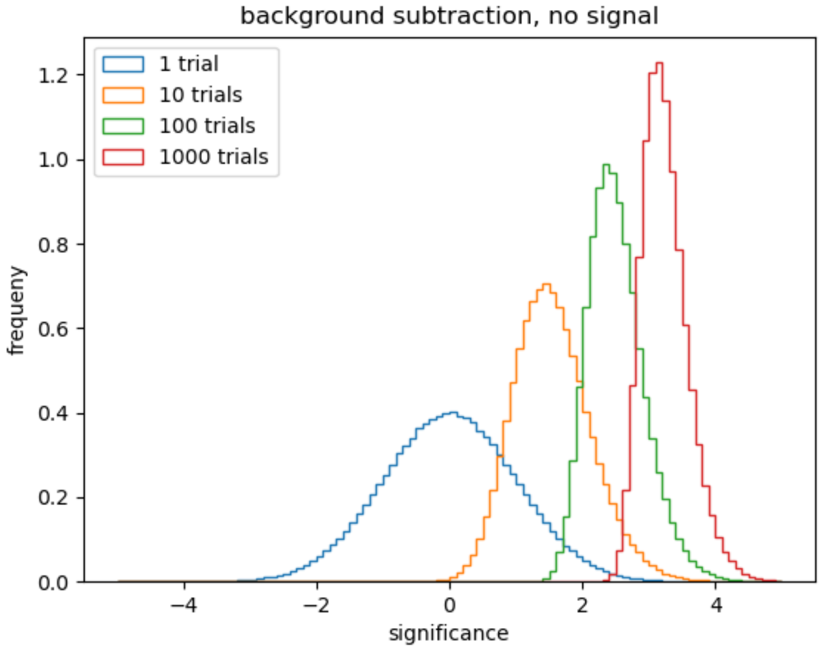
- **This value (called CSTAT) is used as test statistic in model comparisons**

Content

- **Error propagation/change of variables**
- **Statistical and systematic errors**
- **Binned maximum likelihood and model testing**
- **Trial factors /look-elsewhere effect**

Trials





- **Signal searches are often applied repeatedly to several data sets (e.g. transient events) or in many locations (slices/bins in energy/mass/space) of the same (fixed) data set**
- **Estimators like the detection significance have to be corrected for the number of trials**
- **Remark: a full judgement (i.e. interpretation) of the result can depend on measurements conducted by others or earlier**

(Naive) Correction for Trials

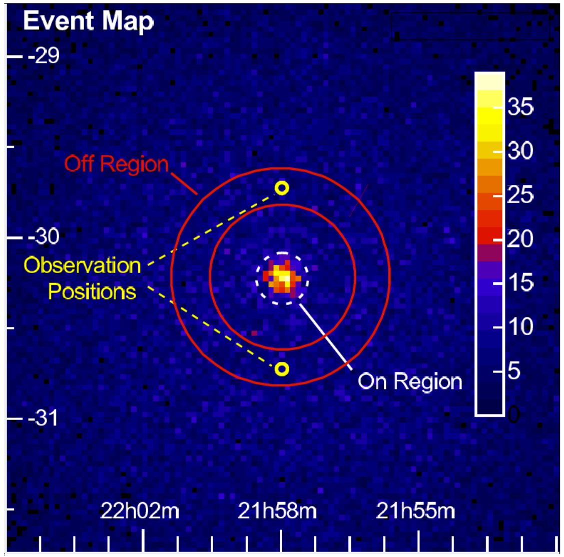
$$\underbrace{(1-p)^N}_{} = 1 - p_N$$

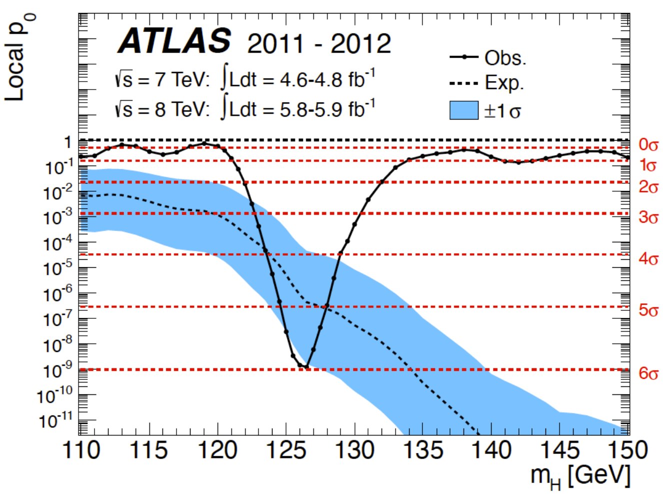
Probality that signal level S was not reached in any of the N trials

$$\Rightarrow p_N = 1 - (1-p)^N$$

- p: probability that some signal level S (e.g. number of events) is reached in a single trial (pre-trial probability)
- p_N: probability that one gets a signal level S after N identical trials (post-trial probability)
- P values can be converted to Gaussian significances as explained in lecture 1
- Remark: Expect numerical problems when evaluating the formula above directly (see appendix for a more stable version)

Complications





- **The naive formula assumed identical trials which is often not the case**
- **Trials are often not independent (e.g. due to overlapping background or signal regions)**
- **Trial factors also occur when an extended parameter space (e.g. mass) is covered**
- **The number of trials N is hard to estimate; one is then usually conservative and avoids underestimating N**
- **MC simulations are straightforward but have computing demands ($O(10^7)$ simulations for 5sigma effect!)**

Thanks

Testing Coverage

Frequentist MC

```
CL = 0.9    //confidence level
N = 1000  //experiments
mu1 = mu2 = 0
for( every possible true mu0 ){
    //test coverage for this mu0
    coverage = 0

    //simulate experiments
    for(i=0;i<N;i=i+1){
        x0 ~ p(x|mu0)
        FreqLimit(CL,x0,mu1,mu2)
        if( mu1<=mu0<=mu2 )
            coverage = coverage + 1
    }
    coverage = coverage/N
    //coverage should equal CL
}
```

Bayesian MC

```
CL = 0.9    //confidence level
N = 1000  //attempts
mu1 = mu2 = 0
for( every possible x0 ){
    //test coverage for this x0
    coverage = 0
    BayesLimit(CL,x0,mu1,mu2)
    //sample posterior
    for(i=0;i<N;i=i+1){
        mu0 ~ p(mu|x0)p(mu)/p(x0)

        if( mu1<=mu0<=mu2 )
            coverage = coverage + 1
    }
    coverage = coverage/N
    //coverage should equal CL
}
```

x,mu : random variables   x0,mu0 : drawn from PDF

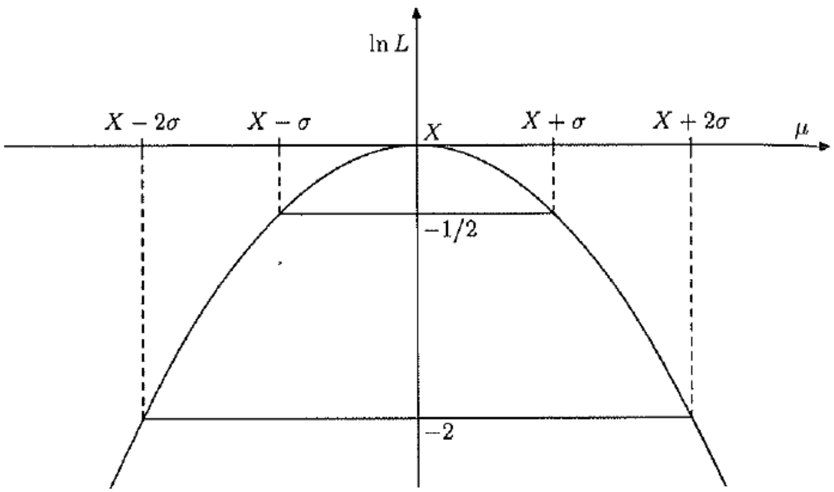
Trial Correction

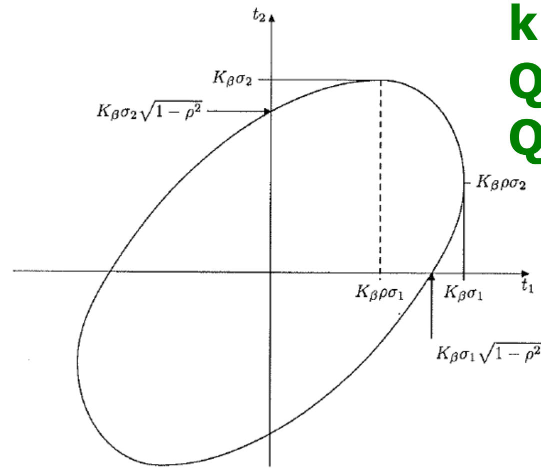
$$
\begin{aligned}
p_n &= 1-(1-p)^n = 1^n-(1-p)^n = (p+(1-p))^n-(1-p)^n \\
&= \sum_{j=0}^{n}\binom{n}{j}p^j(1-p)^{n-j}-(1-p)^n = \sum_{j=1}^{n}\binom{n}{j}p^j(1-p)^{n-j}
\end{aligned}
$$

Approximation for small p: just keep the first two terms in the sum

$$
p_n = np(1-p)^{n-1}+\frac{n(n-1)}{2}p^2(1-p)^{n-2}
$$

Likelihood-based CI





k=2:
Q<=1 (39.3%)
Q<=2.3 (68.3%)

Correlation coefficient ρ=0.5

- **$\ln L(\mu)=\ln L(\mu_{max})-1/2$ for 1 parameter**
- **$\ln L(\mu_1, .., \mu_k)=\ln L(\mu_{max,1}, .., \mu_{max,k}) - 1/2\ F(k,CL)$**
- **F(k,CL) is a constant factor**
- **Integral from 0 to F(k,CL) over a χ^2-distribution with k degrees of freedom gives CL**
  - **F(1,68.3%)=1**
  - **F(2,39.3%)=1, F(2,68.3%)=2.3**
  - **F(3,68.3%)=3.53**

Multivariate Gaussian

$$f(\mathbf{X}) = \frac{1}{(2\pi)^{k/2} |\underset{\sim}{V}|^{\frac{1}{2}}} \exp\left[-\frac{1}{2} (\mathbf{X} - \mu)^{\mathrm{T}} \underset{\sim}{V}^{-1} (\mathbf{X} - \mu) \right]$$

$$\underset{\sim}{V} = \begin{pmatrix} \sigma_1^2 & \rho_{12}\sigma_1\sigma_2 & \cdots & & \rho_{1N}\sigma_1\sigma_N \\ \rho_{12}\sigma_1\sigma_2 & \sigma_2^2 & & & \vdots \\ \vdots & & \ddots & & \\ \vdots & & & \ddots & \vdots \\ \rho_{1N}\sigma_1\sigma_N & \cdots & \cdots & \cdots & \sigma_N^2 \end{pmatrix}$$

- **k Gaussian random variables**
- **Vector of RV X = (x_1, … ,x_k)**
- **Vector of means μ = (μ_1, … ,μ_k)**
- **Correlated unless covariance matrix V is diagonal**
- **V has k(k-1)/2 (off diagonal) + k (diagonal) = k(k+1)/2 independent parameters**
- **V is positive definite**
- **Bell-shaped in k dimensions**

Multivariate Gaussian in 2D

$$f(X,Y) = \frac{1}{2\pi\sigma_X\sigma_Y\sqrt{(1-\rho^2)}}$$

$$\times \exp\left[-\frac{1}{2(1-\rho^2)}\left\{\frac{(X-\mu_X)^2}{\sigma_X{}^2} - 2\rho\frac{(X-\mu_X)(Y-\mu_Y)}{\sigma_X\sigma_Y} + \frac{(Y-\mu_Y)^2}{\sigma_Y{}^2}\right\}\right]$$

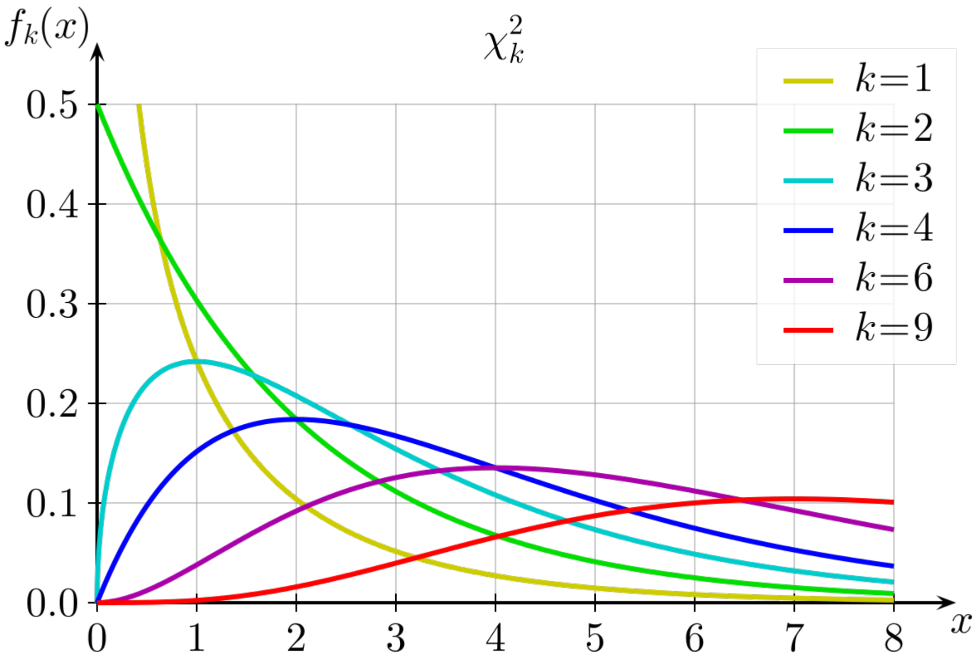
$$\underset{\sim}{V} = \begin{pmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 \end{pmatrix}$$

- **k=2 Gaussian random variables X and Y**
- **2 means μ_x and μ_k**
- **2x2 covariance matrix (3 parameters)**

Covariance Form

$$f(\mathbf{X}) = \frac{1}{(2\pi)^{k/2} |\underset{\sim}{V}|^{\frac{1}{2}}} \exp\left[-\frac{1}{2} (\mathbf{X}-\mu)^T \underset{\sim}{V}^{-1} (\mathbf{X}-\mu) \right]$$

Covariance form Q



- **Contours of constant probability are given by Q=constant**
- **Q is distributed as χ^2(k), (independent of μ!)**
- **Can estimate the probability content of a hyperellipsoid by integrating over the χ^2(k) distribution**