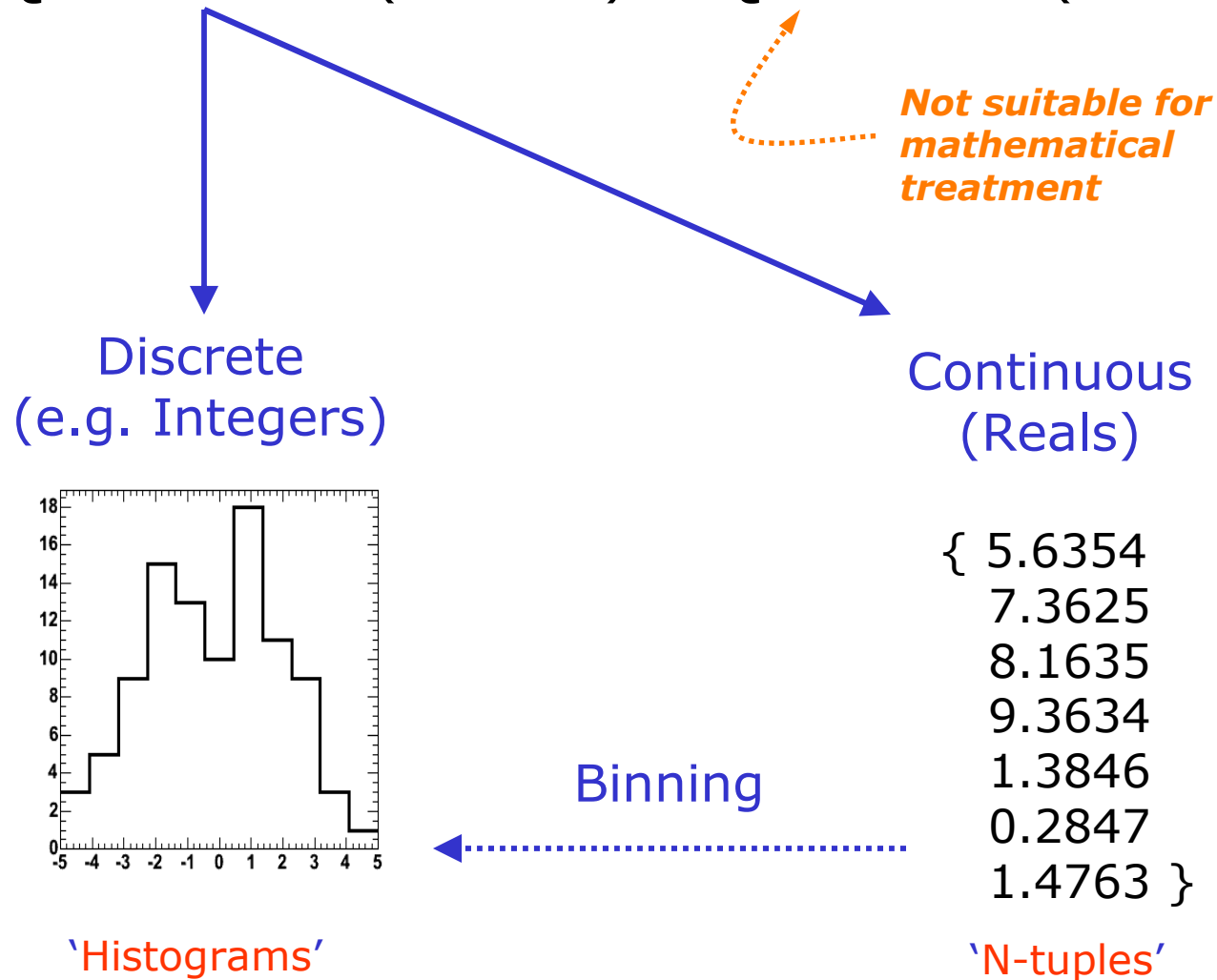


Data – types of data

- **Quantitative** (numeric) vs **Qualitative** (non-numeric)



Credit for these slides:

Wouter Verkerke (NIKHEF)

Describing your data – the Average

- Given a set of *unbinned* data (measurements)

$$\{ \mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_N \}$$

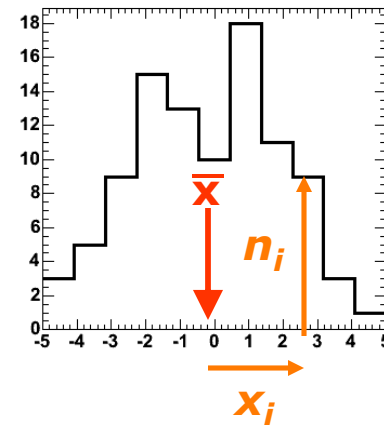
then the mean value of x is

$$\bar{x} = \frac{1}{N} \sum_{i=1}^N x_i$$

- For *binned* data

$$\bar{x} = \frac{1}{N} \sum_{i=1}^N n_i x_i$$

- where n_i is bin count and x_i is bin center
- Unbinned average more accurate due to rounding



Describing your data – Spread

- **Variance $V(x)$** of x expresses how much \mathbf{x} is liable to vary from its mean value \bar{x}

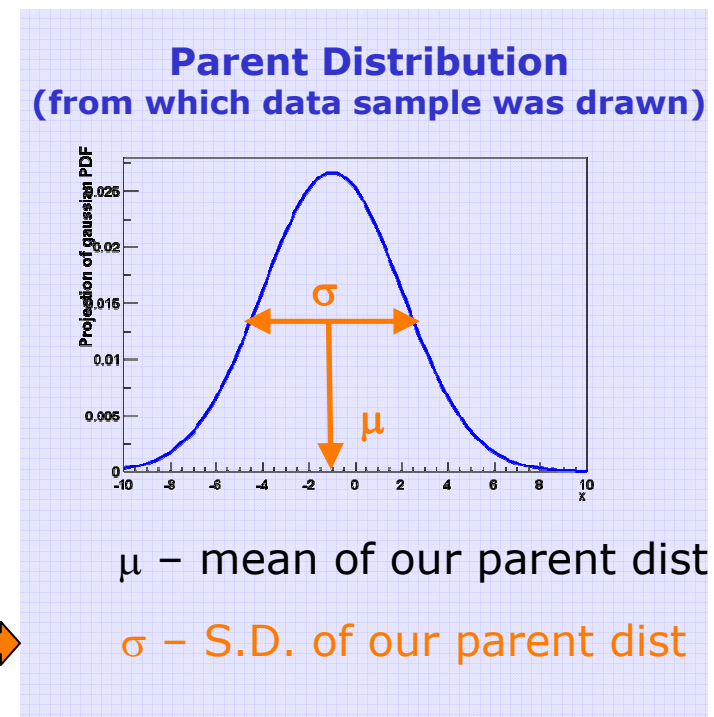
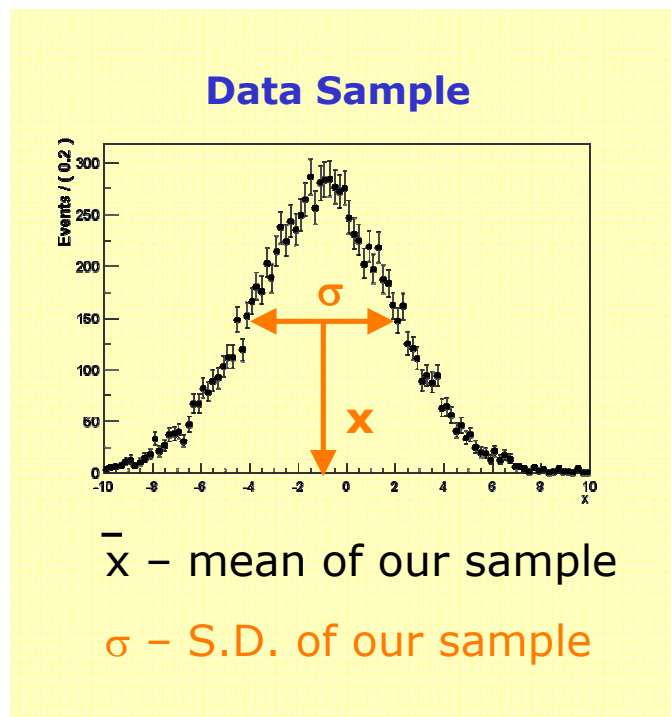
$$\begin{aligned} V(x) &= \frac{1}{N} \sum_i (x_i - \bar{x})^2 \\ &= \frac{1}{N} \sum_i (x_i^2 - 2x_i\bar{x} + \bar{x}^2) \\ &= \frac{1}{N} \sum_i x_i^2 - \frac{1}{N} 2\bar{x} \sum_i x_i + \frac{1}{N} \bar{x}^2 \sum_i 1) \\ &= \overline{x^2} - 2\bar{x}^2 + \bar{x}^2 \\ &= \overline{x^2} - \bar{x}^2 \end{aligned}$$

- **Standard deviation** $\sigma \equiv \sqrt{V(x)} = \sqrt{\overline{x^2} - \bar{x}^2}$

Different definitions of the Standard Deviation

$$\sigma = \sqrt{\frac{1}{N} \sum_i (x_i^2 - \bar{x})^2}$$
 is the S.D. of the **data sample**

- Presumably our data was taken from a parent distributions which has mean μ and S.F. σ

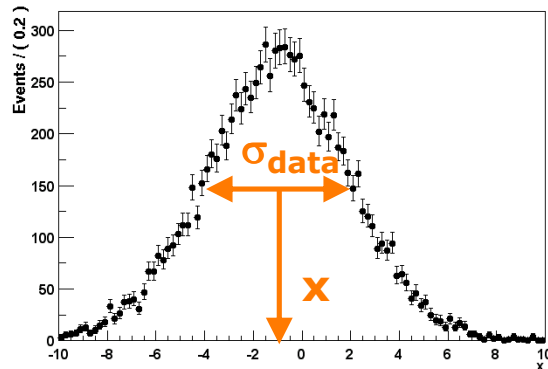


Beware Notational Confusion!

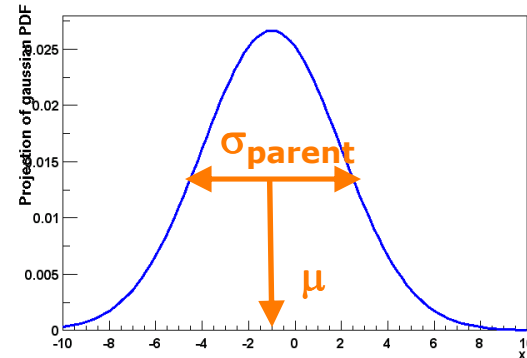
Different definitions of the Standard Deviation

- Which definition of σ you use, σ_{data} or σ_{parent} , is matter of preference, but be clear which one you mean!

Data Sample



Parent Distribution
(from which data sample was drawn)



- In addition, you can get an **unbiased estimate of σ_{parent}** from a given data sample using

$$\hat{\sigma}_{\text{parent}} = \sqrt{\frac{1}{N-1} \sum_i (x^2 - \bar{x})^2} = \hat{\sigma}_{\text{data}} \sqrt{\frac{N}{N-1}}$$

$$\left(\sigma_{\text{data}} = \sqrt{\frac{1}{N} \sum_i (x^2 - \bar{x})^2} \right)$$

More than one variable

- Given *2 variables* x, y and a dataset consisting of pairs of numbers

$$\{ (x_1, y_1), (x_2, y_2), \dots (x_N, y_N) \}$$

- Definition of $\bar{x}, \bar{y}, \sigma_x, \sigma_y$ as usual
- In addition, any *dependence between* x, y described by the **covariance**

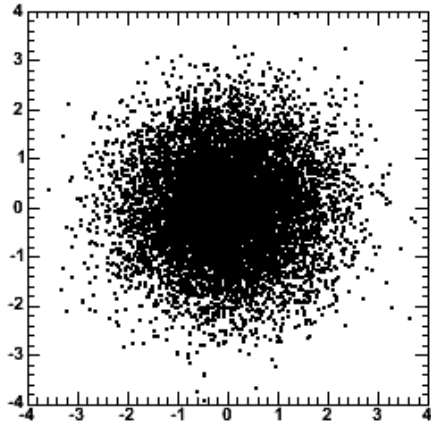
$$\begin{aligned} \text{cov}(x, y) &= \frac{1}{N} \sum_i (x_i - \bar{x})(y_i - \bar{y}) \\ &= \overline{(x - \bar{x})(y - \bar{y})} \\ &= \overline{xy} - \bar{x} \bar{y} \end{aligned}$$

(has dimension $D(x)D(y)$)

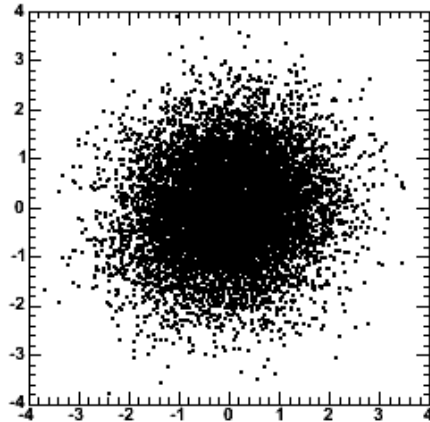
- The dimensionless **correlation coefficient** is defined as $\rho = \frac{\text{cov}(x, y)}{\sigma_x \sigma_y} \in [-1, +1]$

Visualization of correlation

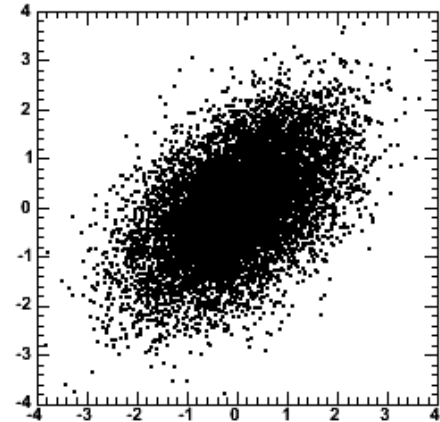
$$\rho = 0$$



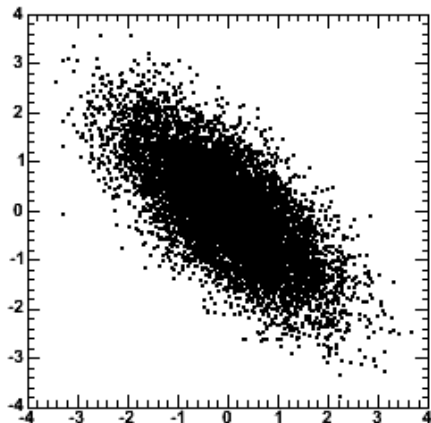
$$\rho = 0.1$$



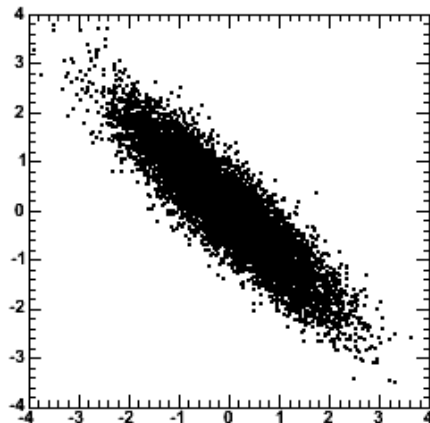
$$\rho = 0.5$$



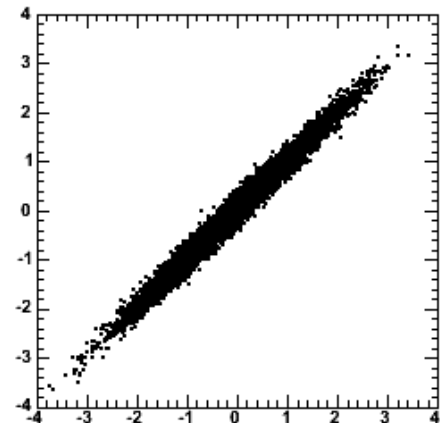
$$\rho = -0.7$$



$$\rho = -0.9$$



$$\rho = 0.99$$



Correlation & covariance in >2 variables

- Concept of covariance, correlation is easily extended to arbitrary number of variables

$$\text{cov}(x_{(i)}, x_{(j)}) = \overline{x_{(i)}x_{(j)}} - \bar{x}_{(i)}\bar{x}_{(j)}$$

- so that $V_{ij} = \text{cov}(x_{(i)}, x_{(j)})$ takes the form of a ***n x n symmetric matrix***
- This is called the ***covariance matrix***, or ***error matrix***
- Similarly the correlation matrix becomes

$$\rho_{ij} = \frac{\text{cov}(x_{(i)}, x_{(j)})}{\sigma_{(i)}\sigma_{(j)}} \longrightarrow V_{ij} = \rho_{ij}\sigma_i\sigma_j$$

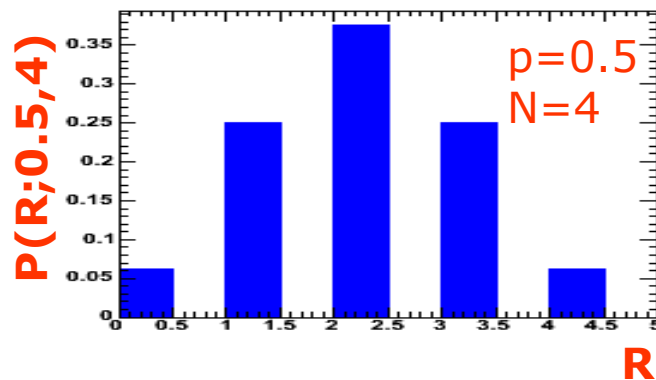
Basic Distributions – The binomial distribution

- Simple experiment – Drawing marbles from a bowl
 - Bowl with marbles, fraction p are black, others are white
 - Draw N marbles from bowl, *put marble back after each drawing*
 - Distribution of R black marbles in drawn sample:

Probability of a
specific outcome
e.g. 'BBBWBWW'

Number of equivalent
permutations for that
outcome

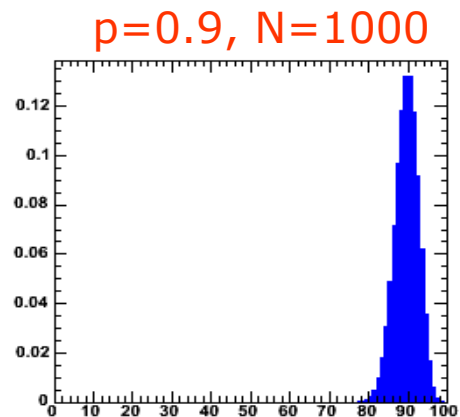
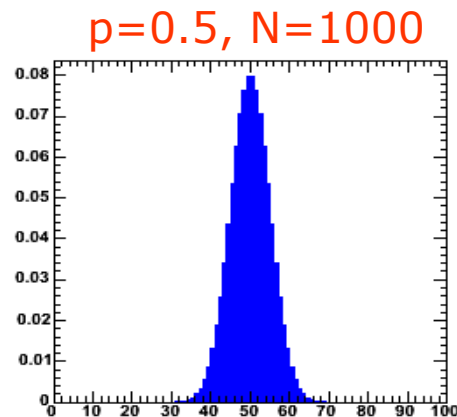
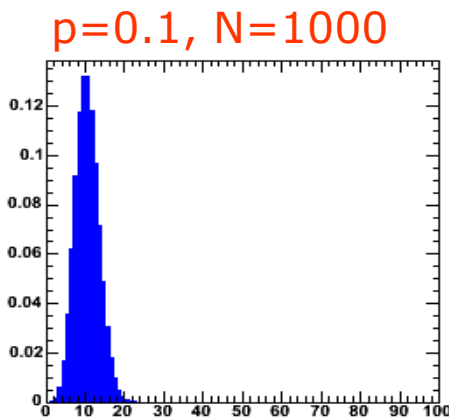
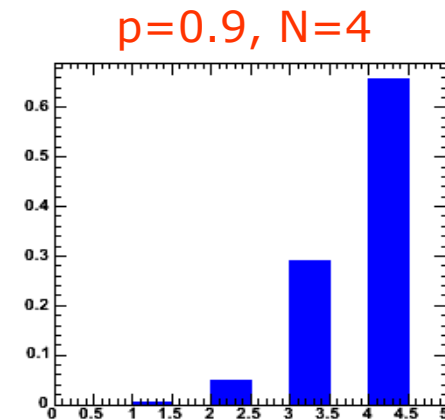
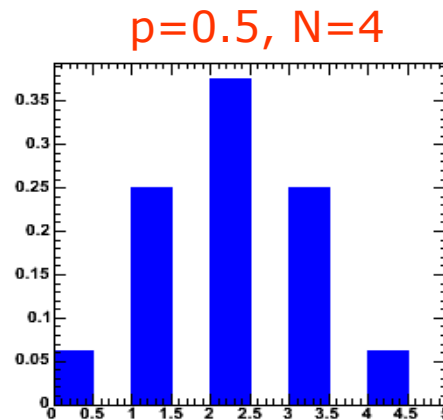
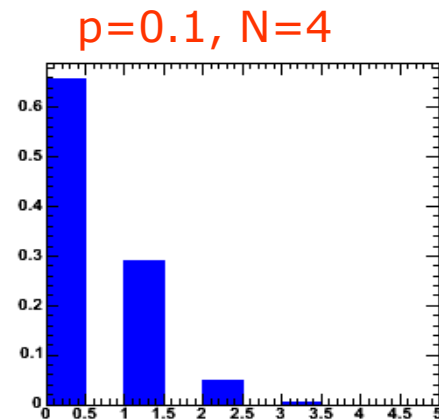
$$P(R; p, N) = p^R (1 - p)^{N-R} \frac{N!}{R!(N - R)!}$$



Binomial distribution


Properties of the binomial distribution

- Mean: $\langle r \rangle = n \cdot p$
- Variance: $V(r) = np(1-p) \Rightarrow \sigma = \sqrt{np(1-p)}$



Basic Distributions – the Poisson distribution

- Sometimes we don't know the equivalent of the number of drawings
 - **Example: Geiger counter**
 - Sharp events occurring in a (time) continuum
- What distribution do we expect in measurement over fixed amount of time?
 - Divide time interval λ in n finite chunks,
 - Take binomial formula with $p=\lambda/n$ and let $n \rightarrow \infty$

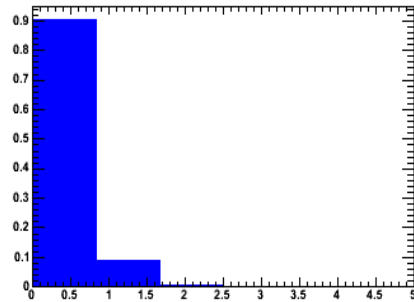
$$P(r; \lambda / n, n) = \frac{\lambda^r}{n^r} \left(1 - \frac{\lambda}{n}\right)^{n-r} \frac{n!}{r!(n-r)!}$$

$$\lim_{n \rightarrow \infty} \frac{n!}{r!(n-r)!} = n^r,$$
$$\lim_{n \rightarrow \infty} \left(1 - \frac{\lambda}{n}\right)^{n-r} = e^{-\lambda}$$

$$P(r; \lambda) = \frac{e^{-\lambda} \lambda^r}{r!}$$

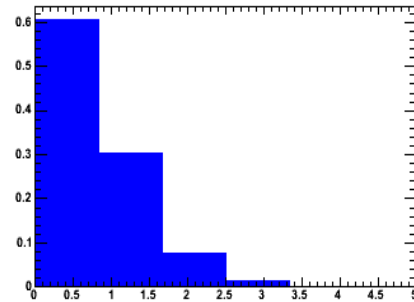
← **Poisson distribution**

Properties of the Poisson distribution

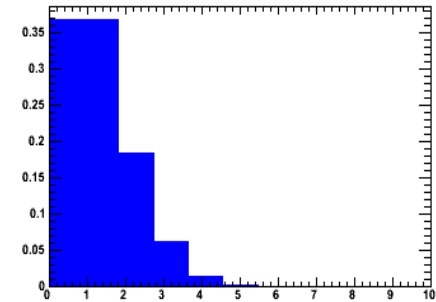
$\lambda=0.1$



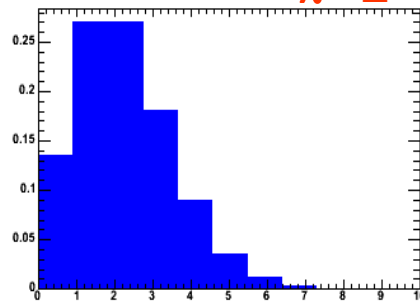
$\lambda=0.5$



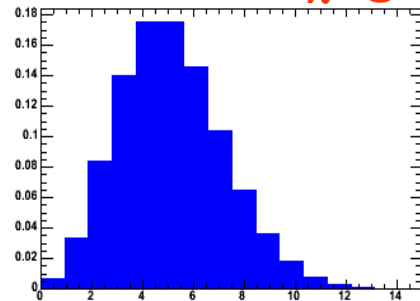
$\lambda=1$



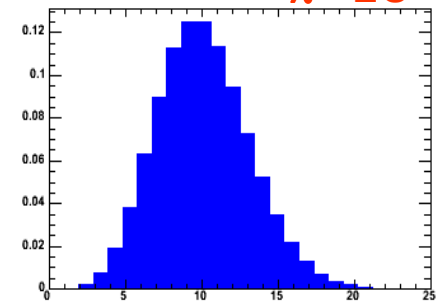
$\lambda=2$



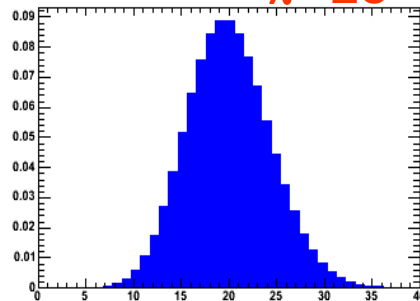
$\lambda=5$



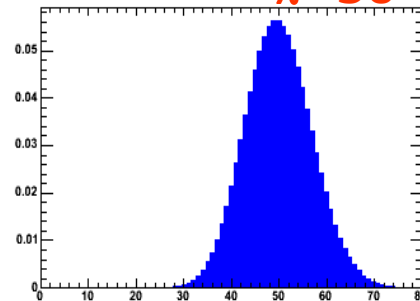
$\lambda=10$



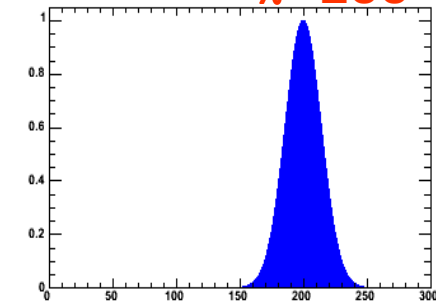
$\lambda=20$



$\lambda=50$



$\lambda=200$



More properties of the Poisson distribution $P(r; \lambda) = \frac{e^{-\lambda} \lambda^r}{r!}$

- Mean, variance: $\langle r \rangle = \lambda$

$$V(r) = \lambda \quad \Rightarrow \quad \sigma = \sqrt{\lambda}$$

- Convolution of 2 Poisson distributions is also a Poisson distribution with $\lambda_{ab} = \lambda_a + \lambda_b$

$$\begin{aligned} P(r) &= \sum_{r_A=0}^r P(r_A; \lambda_A) P(r - r_A; \lambda_B) \\ &= e^{-\lambda_A} e^{-\lambda_B} \sum_{r_A=0}^r \frac{\lambda_A^{r_A} \lambda_B^{r-r_A}}{r_A! (r - r_A)!} \\ &= e^{-(\lambda_A + \lambda_B)} \frac{(\lambda_A + \lambda_B)^r}{r!} \sum_{r_A=0}^r \frac{r!}{(r - r_A)!} \left(\frac{\lambda_A}{\lambda_A + \lambda_B} \right)^{r_A} \left(\frac{\lambda_B}{\lambda_A + \lambda_B} \right)^{r-r_A} \\ &= e^{-(\lambda_A + \lambda_B)} \frac{(\lambda_A + \lambda_B)^r}{r!} \left(\frac{\lambda_A}{\lambda_A + \lambda_B} + \frac{\lambda_B}{\lambda_A + \lambda_B} \right)^r \\ &= e^{-(\lambda_A + \lambda_B)} \frac{(\lambda_A + \lambda_B)^r}{r!} \end{aligned}$$

Basic Distributions – The Gaussian distribution

- Look at *Poisson distribution* in limit of *large N*

$$P(r; \lambda) = e^{-\lambda} \frac{\lambda^r}{r!}$$

Take log, substitute, $r = \lambda + x$,
and use $\ln(r!) \approx r \ln r - r + \ln \sqrt{2\pi r}$

$$\begin{aligned} \ln(P(r; \lambda)) &= -\lambda + r \ln \lambda - (r \ln r - r) - \ln \sqrt{2\pi r} \\ &= -\lambda + r \left[\ln \lambda - \ln \left(\lambda \left(1 + \frac{x}{\lambda} \right) \right) \right] + (\lambda + x) - \ln \sqrt{2\pi \lambda} \\ &\approx x - (\lambda - x) \left(\frac{x}{\lambda} + \frac{x^2}{2\lambda^2} \right) - \ln(2\pi \lambda) \end{aligned}$$

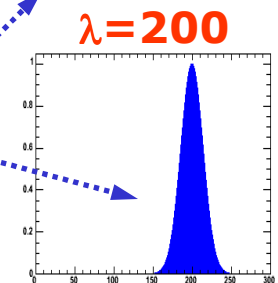
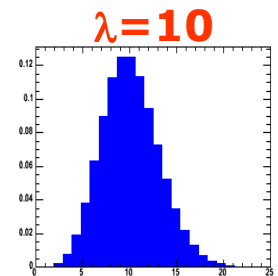
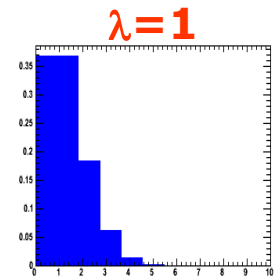
$\ln(1+z) \approx z - z^2/2$

$$\approx \frac{-x^2}{2\lambda} - \ln(2\pi \lambda)$$

Take exp

$$P(x) = \frac{e^{-x^2/2\lambda}}{\sqrt{2\pi \lambda}}$$

Familiar Gaussian distribution,
(approximation reasonable for $N > 10$)

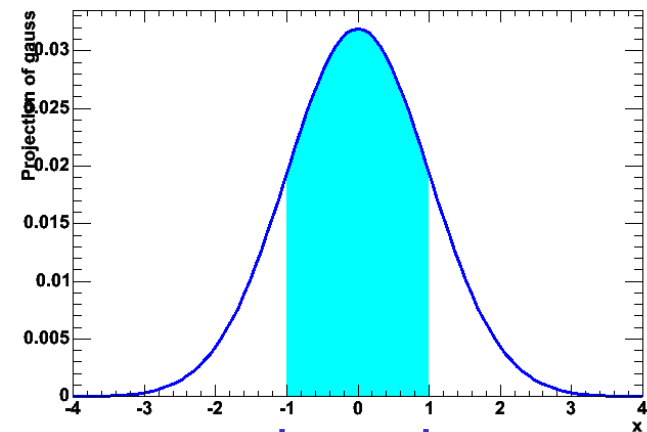


Properties of the Gaussian distribution

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

- *Mean* and *Variance*

$$\begin{aligned}\langle x \rangle &= \int_{-\infty}^{+\infty} x P(x; \mu, \sigma) dx = \mu \\ V(x) &= \int_{-\infty}^{+\infty} (x - \mu)^2 P(x; \mu, \sigma) dx = \sigma^2 \\ \sigma &= \sigma\end{aligned}$$



- Integrals of Gaussian

68.27% within 1σ	$90\% \rightarrow 1.645\sigma$
95.43% within 2σ	$95\% \rightarrow 1.96\sigma$
99.73% within 3σ	$99\% \rightarrow 2.58\sigma$
	$99.9\% \rightarrow 3.29\sigma$

Errors

- Doing an experiment → making measurements
- Measurements not perfect → imperfection quantified in resolution or error
- Common language to quote errors
 - Gaussian standard deviation = $\sqrt{V(x)}$
 - 68% probability that true values is within quoted errors

[NB: 68% interpretation relies strictly on Gaussian sampling distribution, which is not always the case, more on this later]
- Errors are usually Gaussian if they quantify a result that is based on many independent measurements

The Gaussian as 'Normal distribution'

- Why are errors usually Gaussian?
- The **Central Limit Theorem** says
 - If you take the sum X of N independent measurements x_i , each taken from a distribution of mean m_i , a variance $V_i = \sigma_i^2$, the distribution for x

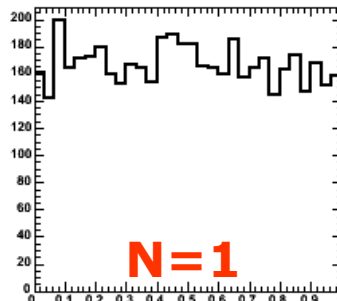
(a) has expectation value $\langle X \rangle = \sum_i \mu_i$

(b) has variance $V(X) = \sum_i V_i = \sum_i \sigma_i^2$

(c) becomes Gaussian as $N \rightarrow \infty$

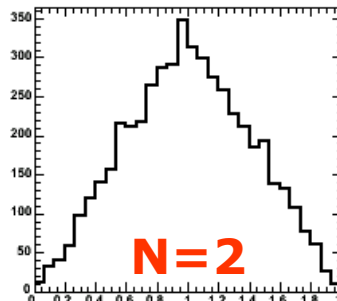
- *Small print: tails converge very slowly in CLT, be careful in assuming Gaussian shape beyond 2σ*

Demonstration of Central Limit Theorem



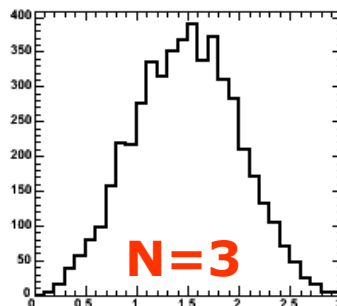
← 5000 numbers taken at random from a uniform distribution between $[0,1]$.

– Mean = $1/2$, Variance = $1/12$

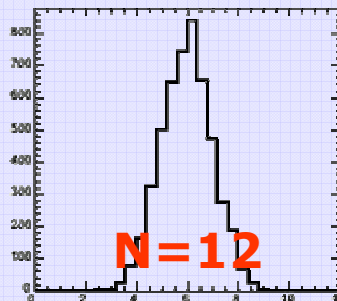


← 5000 numbers, each the sum of 2 random numbers, i.e. $X = x_1 + x_2$.

– Triangular shape



← Same for 3 numbers,
 $X = x_1 + x_2 + x_3$



← Same for 12 numbers, overlaid curve is exact Gaussian distribution

Error propagation – one variable

- Suppose we have $f(x) = ax + b$
- How do you calculate $V(f)$ from $V(x)$?

$$\begin{aligned} V(f) &= \langle f^2 \rangle - \langle f \rangle^2 \\ &= \langle (ax + b)^2 \rangle - \langle ax + b \rangle^2 \\ &= a^2 \langle x^2 \rangle + 2ab \langle x \rangle + b^2 - a \langle x \rangle^2 - 2ab \langle x \rangle - b^2 \\ &= a^2 (\langle x^2 \rangle - \langle x \rangle^2) \\ &= a^2 V(x) \end{aligned} \quad \leftarrow \text{i.e. } \sigma_f = |a| \sigma_x$$

- More general: $V(f) = \left(\frac{df}{dx} \right)^2 V(x) \quad ; \quad \sigma_f = \left| \frac{df}{dx} \right| \sigma_x$
 - But only valid if *linear approximation is good in range of error*

Error Propagation – Summing 2 variables

- Consider $f = ax + by + c$

$$V(f) = a^2 \left(\langle x^2 \rangle - \langle x \rangle^2 \right) + b^2 \left(\langle y^2 \rangle - \langle y \rangle^2 \right) + 2ab \left(\langle xy \rangle - \langle x \rangle \langle y \rangle \right)$$
$$= a^2 V(x) + b^2 V(y) + \underline{2ab \operatorname{cov}(x, y)}$$

Familiar 'add errors in quadrature'
only valid in absence of correlations,
i.e. $\operatorname{cov}(x, y) = 0$

- More general

$$V(f) = \left(\frac{df}{dx} \right)^2 V(x) + \left(\frac{df}{dy} \right)^2 V(y) + 2 \left(\frac{df}{dx} \right) \left(\frac{df}{dy} \right) \operatorname{cov}(x, y)$$
$$\sigma_f^2 = \left(\frac{df}{dx} \right)^2 \sigma_x^2 + \left(\frac{df}{dy} \right)^2 \sigma_y^2 + 2 \left(\frac{df}{dx} \right) \left(\frac{df}{dy} \right) \rho \sigma_x \sigma_y$$

But only valid if *linear approximation is good in range of error* **The correlation coefficient ρ [-1,+1] is 0 if x,y uncorrelated**

Error propagation – multiplying, dividing 2 variables

- Now consider $f = x \cdot y$

$$V(f) = y^2 V(x) + x^2 V(y) \quad (\text{math omitted})$$

$$\left(\frac{\sigma_f}{f} \right)^2 = \left(\frac{\sigma_x}{x} \right)^2 + \left(\frac{\sigma_y}{y} \right)^2$$

- Result similar for $f = x / y$

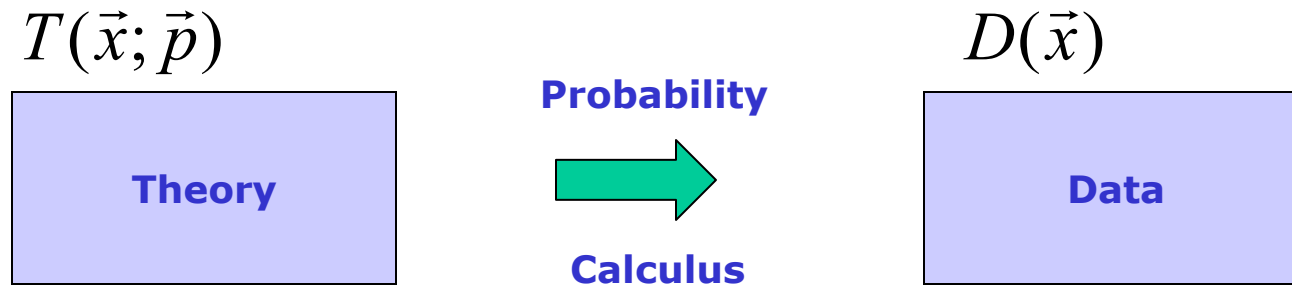
- Other useful formulas

$$\frac{\sigma_{1/x}}{1/x} = \frac{\sigma_x}{x} \quad ; \quad \sigma_{\ln(x)} = \frac{\sigma_x}{x}$$

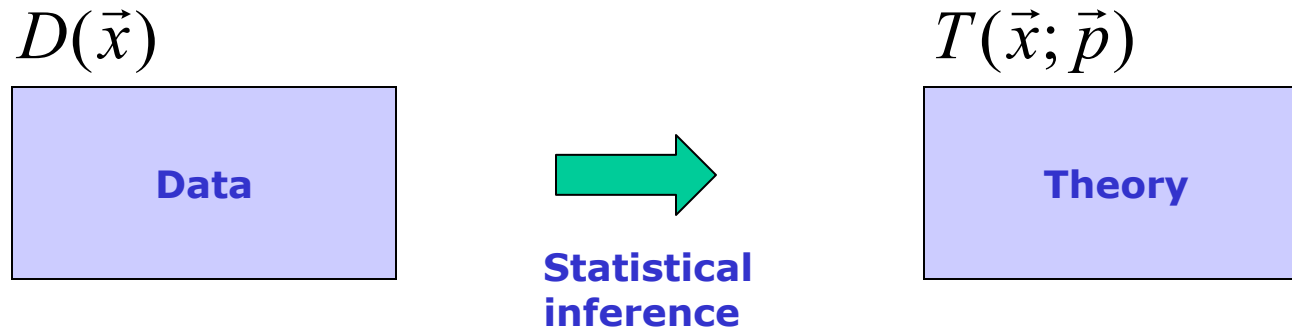
**Relative error on
x, 1/x is the same**

**Error on log is just
fractional error**

Estimation – Introduction



- Given the theoretical distribution parameters p , what can we say about the data

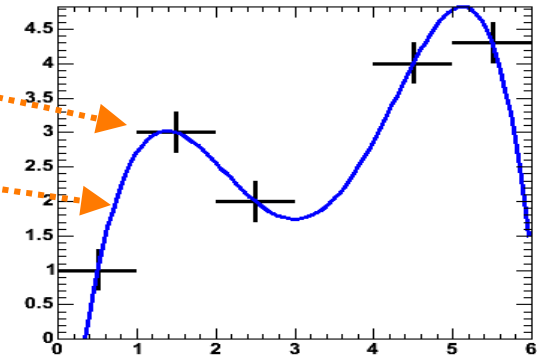


- **Need a procedure to estimate p from D**
 - Common technique – fit!

A well known estimator – the χ^2 fit

- Given a set of points $\{(\vec{x}_i, y_i, \sigma_i)\}$ and a function $f(\mathbf{x}, \mathbf{p})$ define the χ^2

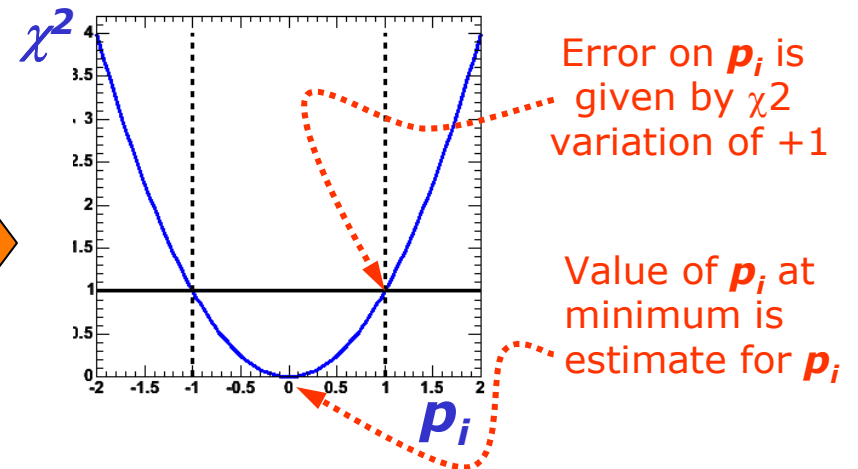
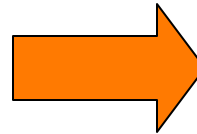
$$\chi^2(\vec{p}) = \sum_i \frac{(y_i - f(\vec{x}; \vec{p}))^2}{\sigma_y^2}$$



- Estimate parameters by minimizing the $\chi^2(\mathbf{p})$ with respect to all parameters p_i

– In practice, look for

$$\frac{d\chi^2(p_i)}{dp_i} = 0$$



- Well known: but why does it work? Is it always right? Does it always give the best possible error?

Basics – What is an estimator?

- An **estimator** is a **procedure** giving a value for a parameter or a property of a distribution as a function of the actual data values, i.e.


$$\hat{\mu}(x) = \frac{1}{N} \sum_i x_i \quad \leftarrow \text{Estimator of the mean}$$

$$\hat{V}(x) = \frac{1}{N} \sum_i (x_i - \bar{\mu})^2 \quad \leftarrow \text{Estimator of the variance}$$

- A perfect estimator is

- **Consistent:** $\lim_{n \rightarrow \infty} (\hat{a}) = a$

- **Unbiased** – *With finite statistics you get the right answer on average*

- **Efficient** $V(\hat{a}) = \langle (\hat{a} - \langle \hat{a} \rangle)^2 \rangle$  This is called the **Minimum Variance Bound**

- ***There are no perfect estimators!***

Likelihood – Another common estimator

- **Definition** of Likelihood

- given $\mathbf{D}(\vec{\mathbf{x}})$ and $\mathbf{F}(\vec{\mathbf{x}}; \vec{\mathbf{p}})$

NB: Functions used in likelihoods must be Probability Density Functions:

$$\int F(\vec{x}; \vec{p}) d\vec{x} \equiv 1, \quad F(\vec{x}; \vec{p}) > 0$$

$$L(\vec{p}) = \prod_i F(\vec{x}_i; \vec{p}), \quad \text{i.e.} \quad L(\vec{p}) = F(x_0; \vec{p}) \cdot F(x_1; \vec{p}) \cdot F(x_2; \vec{p}) \dots$$

- For convenience the **negative log of the Likelihood** is often used

$$-\ln L(\vec{p}) = -\sum_i \ln F(\vec{x}_i; \vec{p})$$

- Parameters are estimated by maximizing the Likelihood, or equivalently minimizing $-\log(L)$

$$\left. \frac{d \ln L(\vec{p})}{d\vec{p}} \right|_{p_i = \hat{p}_i} = 0$$

Variance on ML parameter estimates

- The **estimator** for the **parameter variance** is

$$\hat{\sigma}(p)^2 = \hat{V}(p) = \left(\frac{d^2 \ln L}{d^2 p} \right)^{-1}$$

- I.e. variance is estimated from 2nd derivative of $-\log(L)$ at minimum
- Valid** if estimator is **efficient** and **unbiased**!

From Rao-Cramer-Frechet inequality

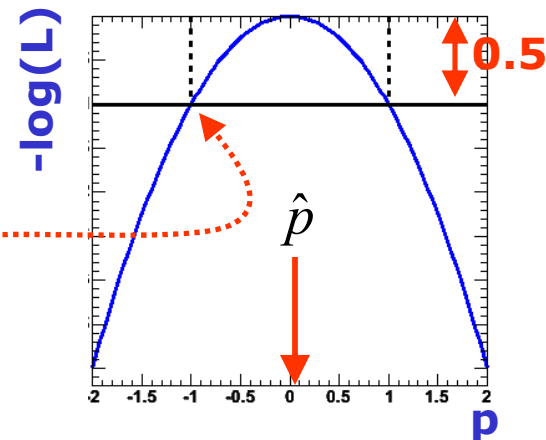
$$V(\hat{p}) \geq 1 + \frac{db}{dp} \bigg/ \left(\frac{d^2 \ln L}{d^2 p} \right)$$

b = bias as function of p , inequality becomes equality in limit of efficient estimator

- Visual interpretation** of variance estimate

- Taylor expand $-\log(L)$ around minimum

$$\begin{aligned} \ln L(p) &= \ln L(\hat{p}) + \frac{d \ln L}{dp} \bigg|_{p=\hat{p}} (p - \hat{p}) + \frac{1}{2} \frac{d^2 \ln L}{d^2 p} \bigg|_{p=\hat{p}} (p - \hat{p})^2 \\ &= \ln L_{\max} + \frac{d^2 \ln L}{d^2 p} \bigg|_{p=\hat{p}} \frac{(p - \hat{p})^2}{2} \\ &= \ln L_{\max} + \frac{(p - \hat{p})^2}{2\hat{\sigma}_p^2} \Rightarrow \ln L(p \pm \sigma) = \ln L_{\max} - \frac{1}{2} \end{aligned}$$



Properties of Maximum Likelihood estimators

- In general, Maximum Likelihood estimators are

- **Consistent** (gives right answer for $N \rightarrow \infty$)
- **Mostly unbiased** (bias $\propto 1/N$, may need to worry at small N)
- **Efficient for large N** (you get the smallest possible error)
- **Invariant:** (a transformation of parameters will Not change your answer, e.g. $(\hat{p})^2 = \widehat{(p^2)}$)

*Use of 2nd derivative of $-\log(L)$
for variance estimate is usually OK*

- MLE efficiency theorem: the MLE will be unbiased and efficient if an unbiased efficient estimator exists
 - Proof not discussed here for brevity
 - Of course this does not guarantee that any MLE is unbiased and efficient for any given problem

More about maximum likelihood estimation

- It's not 'right' it is just sensible
- It does not give you the 'most likely value of p ' – it gives you *the value of p for which this data is most likely*
- Numeric methods are often needed to find the maximum of $\ln(L)$
 - Especially difficult if there is >1 parameter
 - Standard tool in HEP: MINUIT (more about this later)
- Max. Likelihood does **not** give you a **goodness-of-fit** measure
 - If assumed $F(x;p)$ is not capable of describing your data for any p , the procedure will not complain
 - The absolute value of L tells you nothing!

Properties of χ^2 estimators

- Properties of χ^2 estimator follow from properties of ML estimator

$$F(x_i; \vec{p}) = \exp \left[- \left(\frac{y_i - f(x_i; \vec{p})}{\sigma_i} \right)^2 \right]$$

← **Probability Density Function in \vec{p} for single data point $\mathbf{x}_i(\sigma_i)$ and function $f(\mathbf{x}_i; \vec{p})$**



Take log,
Sum over all points \mathbf{x}_i

$$\ln L(\vec{p}) = -\frac{1}{2} \sum_i \left(\frac{y_i - f(x_i; \vec{p})}{\sigma_i} \right)^2 = -\frac{1}{2} \chi^2$$

← **The Likelihood function in \vec{p} for given points $\mathbf{x}_i(\sigma_i)$ and function $f(\mathbf{x}_i; \vec{p})$**

- The χ^2 estimator follows from ML estimator, i.e it is
 - **Efficient, consistent, bias $1/N$, invariant,**
 - **But only in the limit that the error σ_i is truly Gaussian**
 - i.e. need $n_i > 10$ if y_i follows a Poisson distribution
- Bonus: Goodness-of-fit measure – $\chi^2 \approx 1$ per d.o.f

Maximum Likelihood or χ^2 – What should you use?

- χ^2 fit is fastest, easiest
 - Works fine at high statistics
 - Gives absolute goodness-of-fit indication
 - Make (incorrect) Gaussian error assumption on low statistics bins
 - Has bias proportional to $1/N$
 - Misses information with feature size $<$ bin size
- Full Maximum Likelihood estimators most robust
 - No Gaussian assumption made at low statistics
 - No information lost due to binning
 - Gives best error of all methods (especially at low statistics)
 - No intrinsic goodness-of-fit measure, i.e. no way to tell if 'best' is actually 'pretty bad'
 - Has bias proportional to $1/N$
 - Can be computationally expensive for large N
- Binned Maximum Likelihood in between
 - Much faster than full Maximum Likelihood
 - Correct Poisson treatment of low statistics bins
 - Misses information with feature size $<$ bin size
 - Has bias proportional to $1/N$

$$-\ln L(p)_{\text{binned}} = \sum_{\text{bins}} n_{\text{bin}} \ln F(\vec{x}_{\text{bin-center}}; \vec{p})$$

Using weighted data in estimators

- χ^2 fit of histograms with weighted data are straightforward

$$y_i = \sum_i w_i \quad \chi^2 = \sum_i \left(\frac{y_i - f(\vec{x}_i; \vec{p})}{\sigma_i} \right)^2 \quad \sigma_i = \sqrt{\frac{1}{\sum_i w_i^2}}$$

From C.L.T

From C.L.T

- NB: You may no longer be able to interpret $\hat{\sigma}(p) \equiv \sqrt{\hat{V}(p)}$ as a Gaussian error (i.e. 68% contained in 1σ)

- In ML fits implementation of weights easy, but interpretation of errors is not!

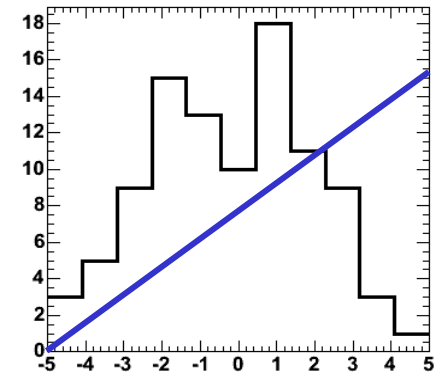
$$-\ln L(\vec{p})_{\text{weighted}} = -\sum_i w_i \ln F(\vec{x}_i; \vec{p})$$

Event weight

- Variance estimate on parameters will be proportional to $\sum_i w_i$
- If $\sum_i w_i < N$ errors will be too small, if $\sum_i w_i > N$ errors will be too large!
- Interpretation of errors from weighted LL fits difficult -- Avoid it if you can

Estimating and interpreting Goodness-Of-Fit

- Fitting determines best set of parameters of given model to describe data
 - Is **'best' good enough?**, i.e.
 - Is it an adequate description, or are there significant and incompatible differences?



'Not good enough'

- Most common test: **the χ^2 test**

$$\chi^2 = \sum_i \left(\frac{y_i - f(\vec{x}_i; \vec{p})}{\sigma_i} \right)^2$$

- If $f(x)$ describes data then $\chi^2 \approx N$, if $\chi^2 \gg N$ something is wrong
- How to quantify meaning of 'large χ^2 '?

How to quantify meaning of 'large χ^2 '

- Probability distr. for χ^2 is given by

$$\chi^2 = \sum_i \left(\frac{y_i - \mu_i}{\sigma_i} \right)^2 \quad \longrightarrow \quad p(\chi^2, N) = \frac{2^{-N/2}}{\Gamma(N/2)} \chi^{N-2} e^{-\chi^2/2}$$

- To make judgement on goodness-of-fit, relevant quantity is integral of above:

$$P(\chi^2; N) = \int_{\chi^2}^{\infty} p(\chi'^2; N) d\chi'^2$$

- **What does χ^2 probability $P(\chi^2, N)$ mean?**

- It is the probability that a function which does genuinely describe the data on N points would give a χ^2 probability as large or larger than the one you already have.
 - Since it is a probability, it is a number in the range [0-1]

Goodness-of-fit – χ^2

- Example for χ^2 probability

- Suppose you have a function **$\mathbf{f}(\mathbf{x};\mathbf{p})$** which gives a χ^2 of 20 for 5 points (histogram bins).
- Not impossible that **$\mathbf{f}(\mathbf{x};\mathbf{p})$** describes data correctly, just unlikely

- How unlikely? $\int_{20}^{\infty} p(\chi^2, 5) d\chi^2 = 0.0012$

- Note: If function has been fitted to the data

- Then you need to account for the fact that parameters have been adjusted to describe the data

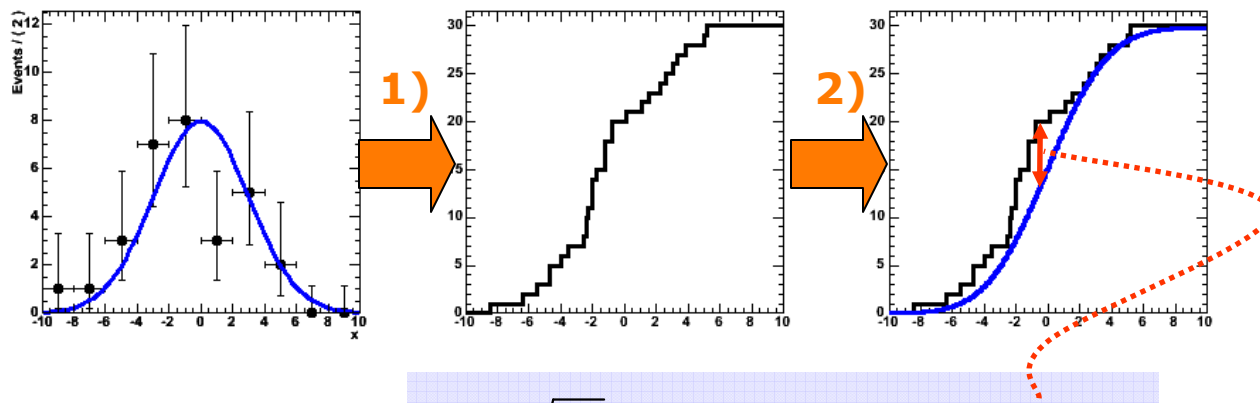
$$N_{\text{d.o.f.}} = N_{\text{data}} - N_{\text{params}}$$

- Practical tips

- To calculate the probability in PAW `'call prob(chi2,ndf)'`
- To calculate the probability in ROOT `'TMath::Prob(chi2,ndf)'`
- For large N, $\sqrt{2\chi^2}$ has a Gaussian distribution with mean $\sqrt{2N-1}$ and $\sigma=1$

Goodness-of-fit – Alternatives to χ^2

- When sample size is very small, it may be difficult to find sensible binning – Look for binning free test
- **Kolmogorov Test**
 - 1) Take all data values, arrange in increasing order and plot cumulative distribution
 - 2) Overlay cumulative probability distribution



– **GOF measure:**

$$d = \sqrt{N} \cdot \max |\text{cum}(x) - \text{cum}(p)|$$

- 'd' large \rightarrow bad agreement; 'd' small – good agreement
- Practical tip: in ROOT: `TH1::KolmogorovTest(TF1&)` calculates probability for you

Systematic errors vs statistical errors

- Definitions

Statistical error = any error in measurement due to statistical fluctuations in data

Systematic errors = **all other errors**

Systematic uncertainty \equiv Systematic error

- But Systematic **error** \neq Systematic **mistake**!

- Suppose we know our measurement needs to be corrected by a factor of 1.05 ± 0.03
- Not correcting the data by factor 1.05 introduces a systematic mistake
- Right thing to do: correct data by factor 1.05 and take uncertainty on factor (0.03) as a systematic error

Source of systematic errors – ‘Good’ and ‘Bad’ errors

- ‘Good’ errors arise from clear causes and can be evaluated
 - Clear cause of error
 - Clear procedure to identify and quantify error
 - Example: Calibration constants,
efficiency corrections from simulation
- ‘Bad’ errors arise from clear causes, but can *not* be evaluated
 - Still clear cause
 - But no unambiguous procedure to quantify uncertainty
 - Example: theory error:
 - Given 2 or more choices of theory model you get 2 or more different answers.
 - What is the error?

Sources of systematic errors – ‘Ugly’ errors

- ‘Ugly’ errors arise from sources that have been overlooked
 - Cause unknown → error unquantifiable
- ‘Ugly’ errors are usually found through failed sanity checks
 - Example: measurement of CP violation on a sample of events that is known to have no CP-violation: You find $A_{CP}=0.10 \pm 0.01$
 - Clearly something is wrong – What to do?
 - 1) **Check your analysis**
 - 2) Check your analysis again
 - 3) Phone a friend
 - 4) Ask the audience
 - ...
 - 99) **Incorporate as systematic error as last and desperate resort!**

What about successful sanity checks?

- Do not incorporate successful checks in your systematic uncertainty
 - Infinite number of successful sanity checks would otherwise lead to infinitely large systematic uncertainty. Clearly not right!
- Define beforehand if a procedure is a sanity check or an evaluation of an uncertainty
 - If outcome of procedure can legitimately be different from zero, it is a systematic uncertainty evaluation
 - If outcome of procedure can only significantly different from zero due to mistake or unknown cause, it is a sanity check

Combining statistical and systematic uncertainty

- Systematic error and statistical error are independent
 - They can be added in quadrature to obtain combined error
 - Nevertheless always quote (also) separately!
 - Also valid procedure if systematic error is not Gaussian: Variances can be added regardless of their shape
 - Combined error usually approximately Gaussian anyway (C.L.T)
- Combining errors a posteriori not only option
 - You can include any systematic error directly in your χ^2 or ML fit:

In χ^2 fit

$$\chi^2 = \chi_{nom}^2 + \left(\frac{p - p_0}{\sigma_p} \right)^2$$

In ML fit

$$-\ln L = - \left[\ln L_{nom} + \frac{1}{2} \left(\frac{p - p_0}{\sigma_p} \right)^2 \right]$$

- Or, for multiple uncertainties with correlations

$$\chi_{pen} = \vec{p}^T V^{-1} \vec{p} \quad ; \quad -\ln L_{pen} = -\frac{1}{2} (\vec{p}^T V^{-1} \vec{p})$$

PAW

(“Physics Analysis Workstation”)

A program for graphical & statistical analysis

- ☞ PAW is an INTERACTIVE SYSTEM
- ☞ PAW provides a set of COMMANDS acting on specific objects
- ☞ The commands structure of PAW is a TREE structure
- ☞ The general structure of the tree is:

OBJECT/ACTION

Example:

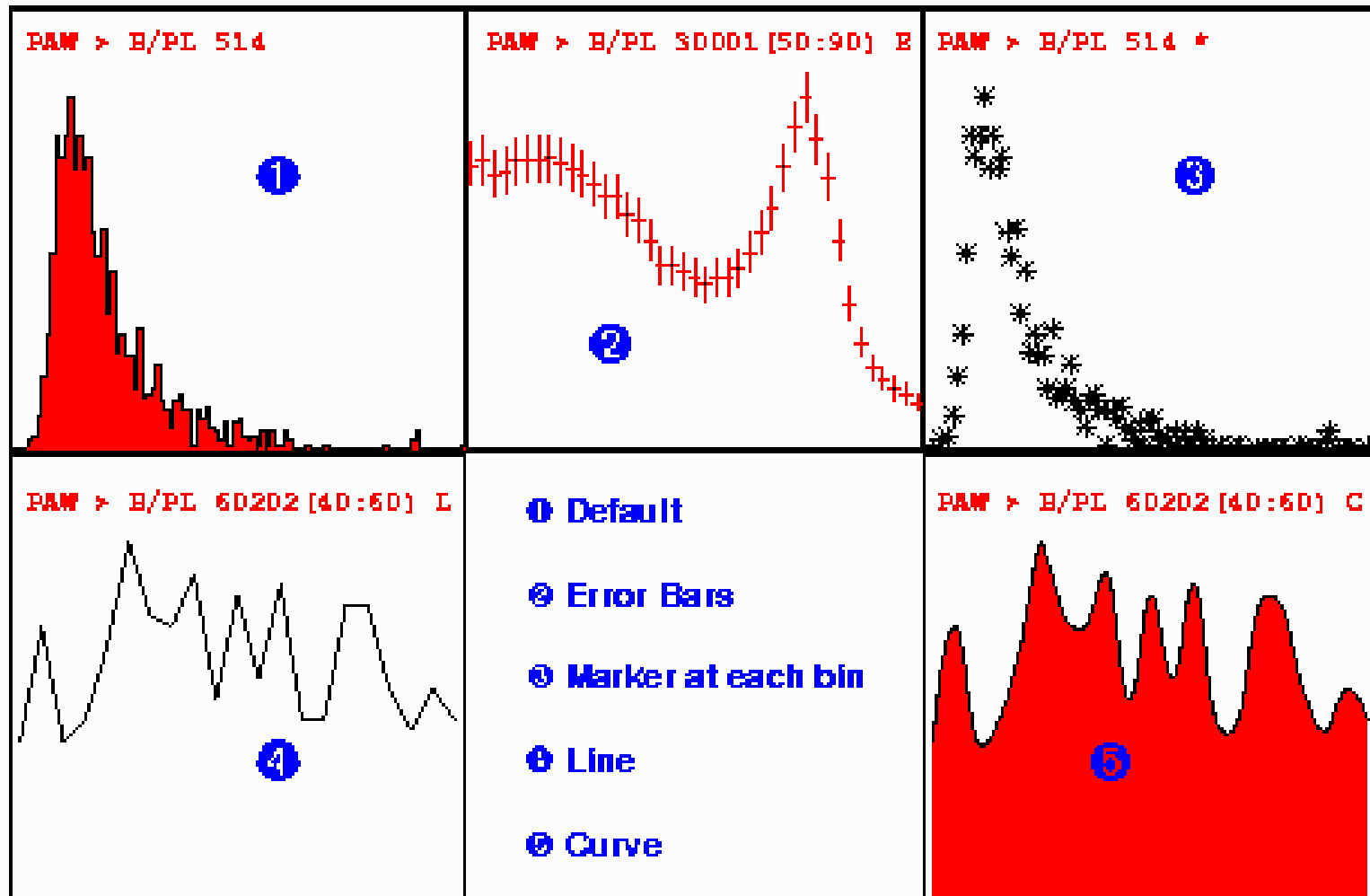
NTUPLE/PLOT

HISTOGRAM/PROJECT

VECTOR/DRAW

- ☞ PAW commands can be grouped into MACROS

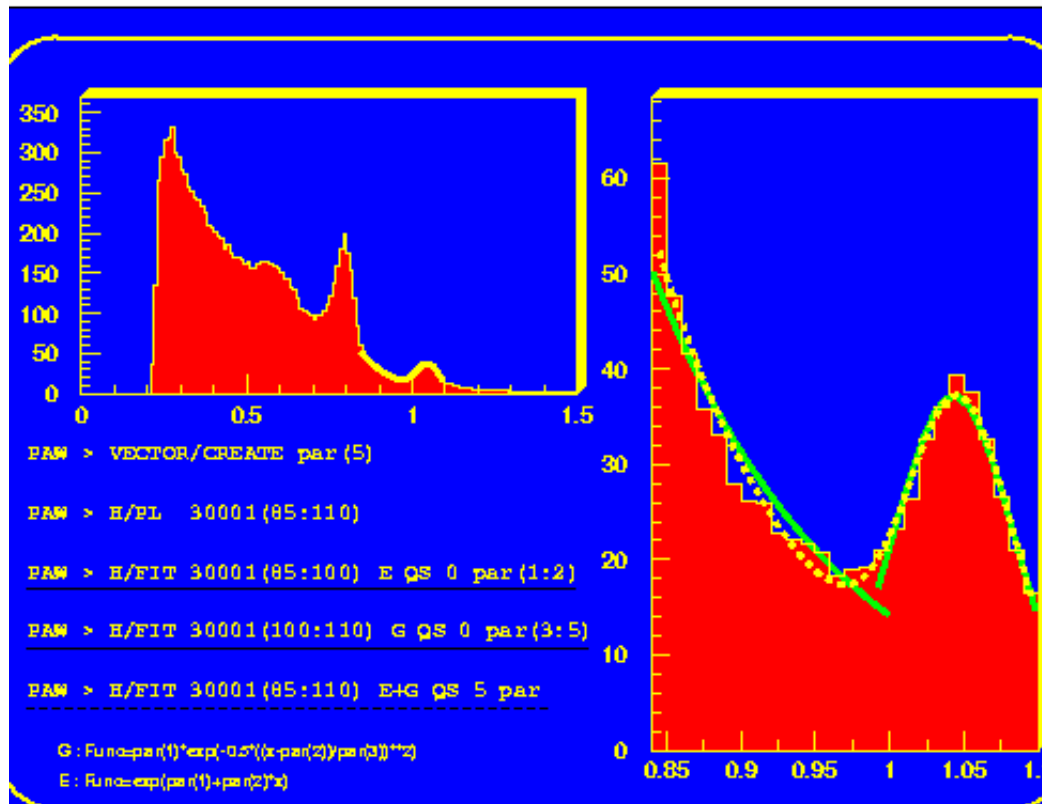
Drawing Histograms in PAW



Fitting Histograms in PAW

The HISTOGRAM/FIT command

```
HISTOGRAM/FIT ID FUNC [ CHOPT NP PAR STEP PMIN PMAX ERRPAR ]
ID          Histogram Identifier
FUNC        Function name
CHOPT       Options
NP          Number of parameters
PAR         Vector of parameters
STEP        Vector of steps size
PMIN        Vector of lower bounds
PMAX        Vector of upper bounds
ERRPAR      Vector of errors on parameters
```



- Automatic
- To show the values of fitted parameters:

PAW > opt stat

(prior to the fit)

The End

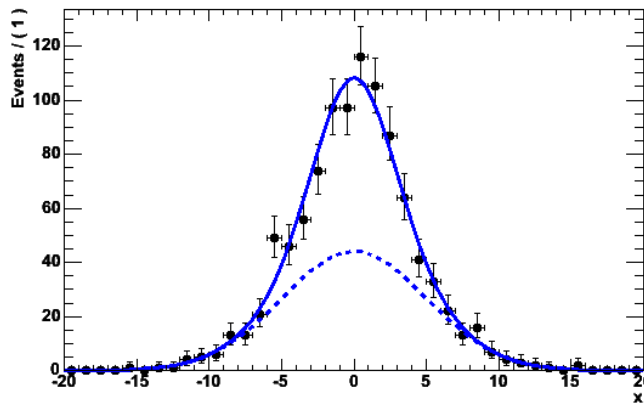
- Some material for further reading
 - R. Barlow, *Statistics: A Guide to the Use of Statistical Methods in the Physical Sciences*, Wiley, 1989
 - L. Lyons, *Statistics for Nuclear and Particle Physics*, Cambridge University Press,
 - G. Cowan, *Statistical Data Analysis*, Clarendon, Oxford, 1998
(See also his 10 hour post-graduate web course:
http://www.pp.rhul.ac.uk/~cowan/stat_course)

Thanks again to Wouter Verkerke (at NIKHEF)
for these slides!

Mitigating fit stability problems

- Strategy I – More orthogonal choice of parameters
 - Example: fitting sum of 2 Gaussians of similar width

$$F(x; f, m, s_1, s_2) = fG_1(x; s_1, m) + (1 - f)G_2(x; s_2, m)$$



HESSE correlation matrix

PARAMETER	CORRELATION COEFFICIENTS				
NO.	GLOBAL	[f]	[m]	[s1]	[s2]
[f]	0.96973	1.000	-0.135	0.918	0.915
[m]	0.14407	-0.135	1.000	-0.144	-0.114
[s1]	0.92762	0.918	-0.144	1.000	0.786
[s2]	0.92486	0.915	-0.114	0.786	1.000

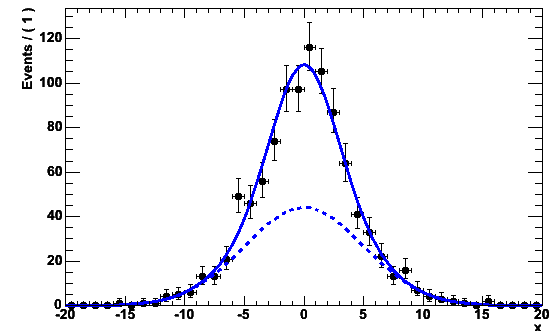
Widths s_1, s_2
strongly correlated
fraction f

Mitigating fit stability problems

- Different parameterization:

$$fG_1(x; s_1, m_1) + (1 - f)G_2(x; \underline{s_1 \cdot s_2}, m_2)$$

PARAMETER	CORRELATION COEFFICIENTS				
NO.	GLOBAL	[f]	[m]	[s1]	[s2]
[f]	0.96951	1.000	-0.134	0.917	-0.681
[m]	0.14312	-0.134	1.000	-0.143	0.127
[s1]	0.98879	0.917	-0.143	1.000	-0.895
[s2]	0.96156	0.681	0.127	-0.895	1.000

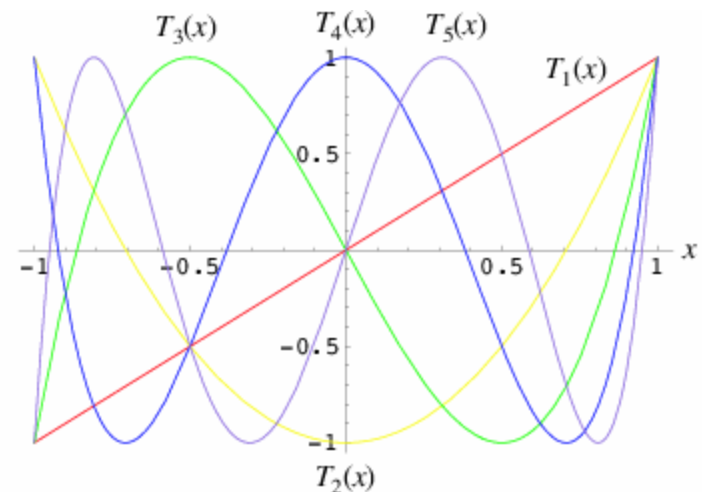
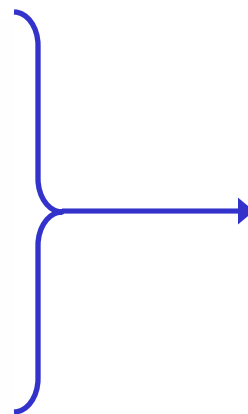


- Correlation of width s2 and fraction f reduced from 0.92 to 0.68
 - Choice of parameterization matters!
- Strategy II – Fix all but one of the correlated parameters
 - If floating parameters are highly correlated, some of them may be redundant and not contribute to additional degrees of freedom in your model

Mitigating fit stability problems -- Polynomials

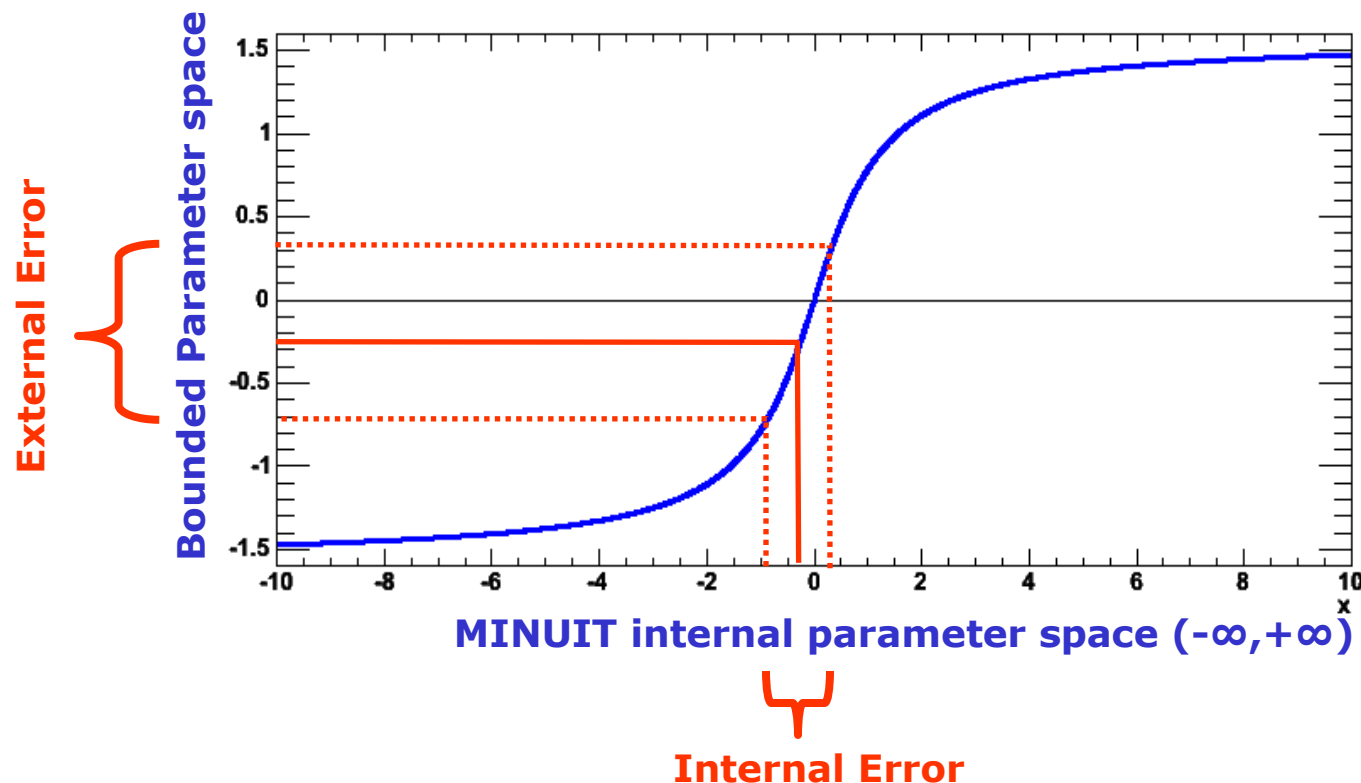
- **Warning:** Regular parameterization of polynomials $a_0 + a_1x + a_2x^2 + a_3x^3$ nearly always results in strong correlations between the coefficients a_i .
 - *Fit stability problems, inability to find right solution common at higher orders*
- **Solution:** Use existing parameterizations of polynomials that have (mostly) uncorrelated variables
 - **Example: Chebychev polynomials**

$$\begin{aligned}T_0(x) &= 1 \\T_1(x) &= x \\T_2(x) &= 2x^2 - 1 \\T_3(x) &= 4x^3 - 3x \\T_4(x) &= 8x^4 - 8x^2 + 1 \\T_5(x) &= 16x^5 - 20x^3 + 5x \\T_6(x) &= 32x^6 - 48x^4 + 18x^2 - 1.\end{aligned}$$



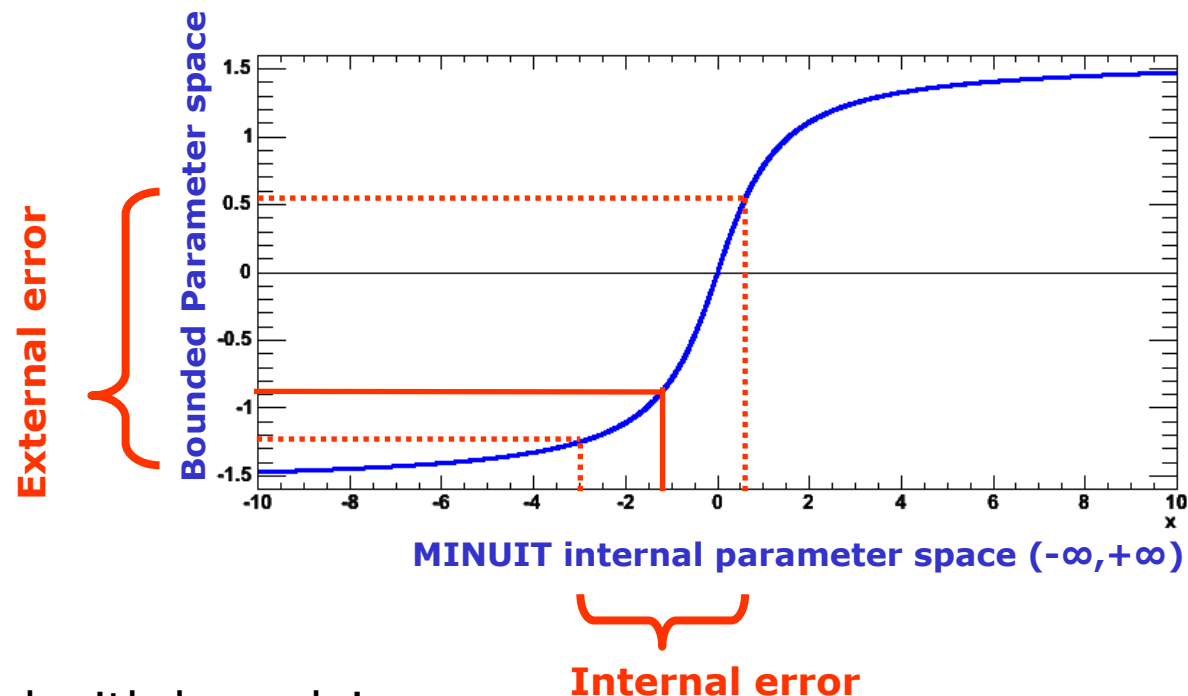
Practical estimation – Bounding fit parameters

- Sometimes it is desirable to bound the allowed range of parameters in a fit
 - Example: a fraction parameter is only defined in the range $[0,1]$
 - MINUIT option 'B' maps finite range parameter to an internal infinite range using an $\arcsin(x)$ transformation:



Practical estimation – Bounding fit parameters

- If fitted parameter values is close to boundary, **errors** will become **asymmetric** (and possible incorrect)



- So be careful with bounds!
 - If boundaries are imposed to avoid region of instability, look into other parameterizations that naturally avoid that region
 - If boundaries are imposed to avoid 'unphysical', but statistically valid results, consider not imposing the limit and dealing with the 'unphysical' interpretation in a later stage

Practical Estimation – Verifying the validity of your fit

- How to validate your fit? – You want to demonstrate that
 - 1) Your fit procedure gives on average the correct answer '**no bias**'
 - 2) The uncertainty quoted by your fit is an accurate measure for the statistical spread in your measurement '**correct error**'
- **Validation is important for low statistics fits**
 - **Correct behavior not obvious a priori due to intrinsic ML bias proportional to $1/N$**
- Basic validation strategy – **A simulation study**
 - 1) Obtain a large sample of simulated events
 - 2) Divide your simulated events in $O(100-1000)$ samples with the same size as the problem under study
 - 3) Repeat fit procedure for each data-sized simulated sample
 - 4) Compare average value of fitted parameter values with generated value → **Demonstrates (absence of) bias**
 - 5) Compare spread in fitted parameters values with quoted parameter error → **Demonstrates (in)correctness of error**

Fit Validation Study – Low statistics example

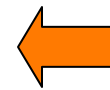
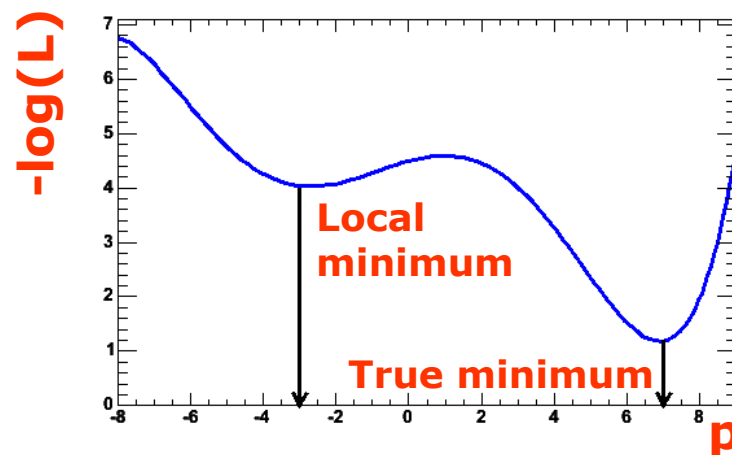
- Special care should be taken when fitting small data samples
 - Also if fitting for small signal component in large sample
- Possible causes of trouble
 - χ^2 estimators may become approximate as Gaussian approximation of Poisson statistics becomes inaccurate
 - ML estimators may no longer be efficient
 - error estimate from 2nd derivative may become inaccurate
 - Bias term proportional to $1/N$ of ML and χ^2 estimators may no longer be small compared to $1/\sqrt{N}$
- In general, absence of bias, correctness of error can not be assumed. How to proceed?
 - Use unbinned ML fits only – most robust at low statistics
 - **Explicitly verify the validity of your fit**

Practical estimation – Numeric χ^2 and $-\log(L)$ minimization

- For most data analysis problems minimization of χ^2 or $-\log(L)$ **cannot be performed analytically**
 - Need to rely on numeric/computational methods
 - In >1 dimension **generally a difficult problem!**
- But no need to worry – Software exists to solve this problem for you:
 - **Function minimization workhorse in HEP many years: MINUIT**
 - MINUIT does function minimization and error analysis
 - It is used in the PAW, ROOT fitting interfaces behind the scenes
 - **It produces a lot of useful information, that is sometimes overlooked**
 - Will look in a bit more detail into MINUIT output and functionality next

Numeric χ^2 / $-\log(L)$ minimization – Proper starting values

- For all but the most trivial scenarios it is not possible to automatically find reasonable starting values of parameters
 - This may come as a disappointment to some...
 - So you need to supply good starting values for your parameters



Reason: There may exist multiple (local) minima in the likelihood or χ^2

- Supplying good initial uncertainties on your parameters helps too
- Reason: Too large error will result in MINUIT coarsely scanning a wide region of parameter space. It may accidentally find a far away local minimum