Data Analysis

Lecture 2: Distributions, estimators

In this lecture

- Distributions
  - Properties
  - Main distributions
- Point (parameter) estimation
  - Maximum likelihood method
  - Least-squares method

Properties of distributions

- **Probability density function** (PDF) = *f(x)*
- **Expectation**
  - Expectation of any random function *g(x)*:

$$E(g) = \int g(x) f(x) dX$$

  - Expectation of x ≡ **mean** of the *f(x)* ≡ **expected value** of x :

$$E(x) = \mu = \bar{x} = \langle x \rangle = \int x f(x) dx$$

- **Variance**

$$V(x) = \sigma^2 = E\left[(x-\mu)^2\right] = E(x^2) - \mu^2 = \int (x-\mu)^2 f(x) dx$$

  - σ is called the **standard deviation**
- *E(x)* is a measure of the **location** of the distribution
- *V(x)* is a mesure of the **spread** of the distribution

Moments

$\mu_n = E(x^n)$	is the nth algebraic moment
$V_n = E\{[x^n - E(x)]^n\}$	is the nth central moment
$\mu'_n = E(\lvert x^n \rvert)$	is the nth absolute moment
$V'_n = E\{\lvert x^n - E(x) \rvert^n\}$	is the nth absolute central moment

- **The coefficient of skewness** $\gamma_1 = \frac{V_3}{V_2^{3/2}}$
  A measure of the skewness of the distribution
- **The coefficient of kurtosis** $\gamma_2 = \frac{V_4}{V_2^2} - 3$
  A measure of the "peakedness" of the distribution

Covariances and correlations

- Joint PDF for two random variables = $f(x,y)$
- The **mean** and the **variance** of x and y:

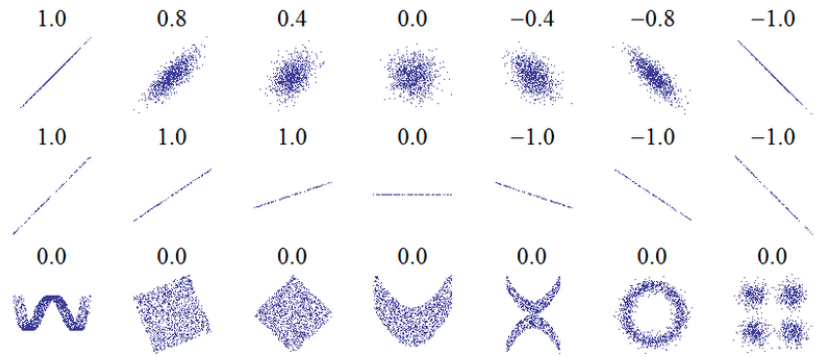
$$\mu_x = E(x) = \iint x f(x,y)dxdy \qquad \mu_y = E(y) = \iint y f(y,y)dxdy$$

$$\sigma_x^2 = E\left[(x-\mu_x)^2\right] \qquad \sigma_y^2 = E\left[(y-\mu_y)^2\right]$$

- **Covariance** $\operatorname{cov}(x,y) = E\left[(x-\mu_x)(y-\mu_y)\right] = E(xy) - E(x) - E(y)$
- **Correlation coefficient** $\operatorname{corr}(x,y) = \rho(x,y) = \rho_{xy} = \dfrac{\operatorname{cov}(x,y)}{\sigma_x \sigma_y}$
- **Covariance/Variance/Error matrix**:

$$\boldsymbol{V} = \begin{bmatrix} \operatorname{cov}(x,x) & \operatorname{cov}(x,y) \\ \operatorname{cov}(x,y) & \operatorname{cov}(y,y) \end{bmatrix} = \begin{bmatrix} \sigma_x^2 & \rho_{xy}\sigma_x\sigma_y \\ \rho_{xy}\sigma_x\sigma_y & \sigma_y^2 \end{bmatrix}$$

Correlations - illustration



Binomial distribution

Variable	r, positive integer $\leq N$
Parameters	N, positive integer; p, $0 \leq p \leq 1$
Probability function	$P(r;N,p) = \binom{N}{r} p^r (1-p)^{N-r}$
Mean	$E(r) = Np$
Variance	$V(r) = Np(1-p)$
Usage example	Example – Z decay: - $p = BR(Z \rightarrow ee) = 3\%$ - $P(5;80,0.03) = 6\%$ probability to find exactly 5 ee events out of 80 Z decays
Comment	$P(r;N,p)$ is a probability of finding exactly r sucesses in N trials, when probability of sucess in each single trial is a constant, p

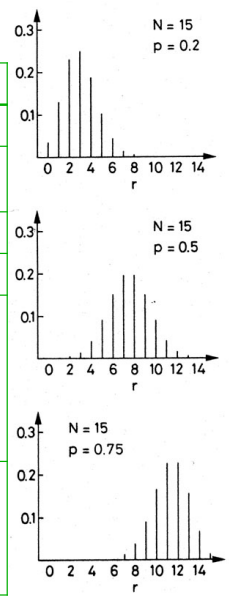


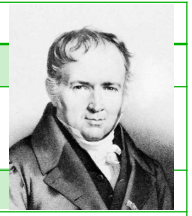
Figure from http://nedwww.ipac.caltech.edu/level5/Leo/Figures/figure1.jpeg

Multinomial distribution

Variable	r_i, $i = 1, \ldots k$, positive integers $\leq$ N
Parameters	N, positive integer k, positive integer $\sum_{i=1}^{k} p_i = 1$ $p_i \geq 0$, $i = 1, \ldots k$,
Probability function	$P(r_1,\ldots,r_k;N,p_1,\ldots,p_k) = \dfrac{N!}{r_1!\cdots r_k!} p_1^{r_1} \cdots p_k^{r_k}$
Mean	$E(r_i) = Np_i$
Variance	$V(r_i) = Np_i(1-p_i)$
Usage example	Histogram containing N events distributed in k bins, with r_i events in the ith bin
Comment	• Multinomial distribution is the generalization of the binomial distribution to the case of more than two possible outcomes of an experiment • When $p_i << 1$ (many bins) $V(r_i) \sim Np_i = r_i$

Poisson distribution

Variable	r, positive integer
Parameters	μ, positive real number
Probability function	$P(r;\mu) = \frac{\mu^r e^{-\mu}}{r!}$
Mean	$E(r) = \mu$
Variance	$V(r) = \mu$
Usage example	Number of events r collected after integrated luminosity $\int Ldt$. Expected number of events is $\mu = \sigma \int Ldt$. σ is the cross section.
Comments	• $P(r;\mu)$ expresses the probability of a number of events occurring in a fixed period of time if these events occur with a known average rate and indepedently of the time since the last event • μ represents expected number of events in a given time interval • Time between two sucessive events is exponentially distributed • Poisson distribution is also called Poissonian



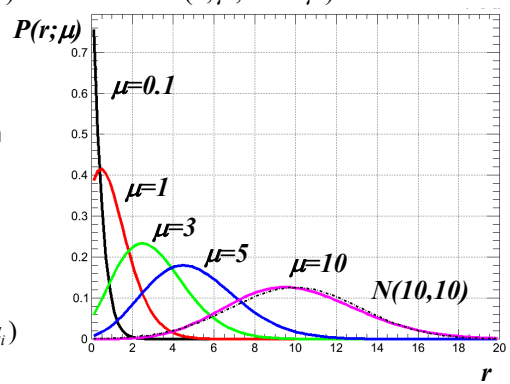
Siméon-Denis Poisson (1781-1840)

Poisson distribution

- For a large μ Poisson distribution converges towards a Gaussian distribution $Pois(r;\mu) \xrightarrow{N>>} Gauss(r;\mu,\sigma^2 = \mu)$
- **Sum of Poisson** distributed random variables also follows a Poisson distribution whose parameter is sum of the component parameters

$$X_i \sim Pois(r;\mu_i)$$

$$Y = \sum_i X_i \sim Pois(r;\sum_i \mu_i)$$

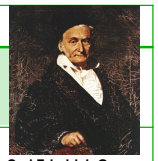


- F.g. When combining signal (s) and background (b)

$$P(r;s,b) \sim Pois(r;s+b)$$

Normal or Gaussian distribution

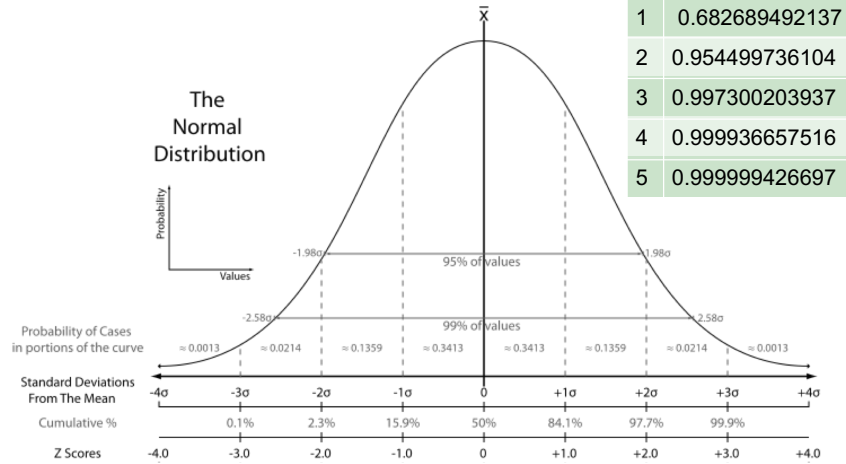
Variable	x, positive real number
Parameters	μ, real number σ, real number
Probability density function	$f(x) = N(\mu,\sigma^2) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left[-\frac{1}{2}\frac{(x-\mu)^2}{\sigma^2}\right]$
Mean	$E(x) = \mu$
Variance	$V(x) = \sigma^2$
Cumulative distribution	$F(x) = \phi\left(\frac{x-\mu}{\sigma}\right); \quad \phi(Z) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{Z} e^{-\frac{1}{2}x^2} dx$
Comments	• The most important distribution in statistics • The half-width at half-height is 1.176σ • $N(0,1)$ is called *standard* Normal density • Any linear combination of the x_i is also Normal



Carl Friedrich Gauss (1777-1855)

Gaussian – some properties

n	Area ±nσ
1	0.682689492137
2	0.954499736104
3	0.997300203937
4	0.999936657516
5	0.999999426697



Why is Gauss Normal?

- **Central limit theorem:**

  If we have a set of N independent variables x_i, each from a distribution with mean μ_i and variance σ_i^2, then the distribution of the sum $X = \Sigma x_i$

  a) has a mean $<X> = \Sigma \mu_i$,

  b) has a variance $V(X) = \Sigma \sigma_i^2$,

  c) becomes Gaussian as $N \rightarrow \infty$.
- Therefore, no matter what the distributions of original variables may have been, their sum will be Gaussian in a large N limit
  - Example: measurements errors
- Example (adopted from Barlow):

  "*Human heights are well described by a Gaussian distribution, as many other anatomical measurements, as these are due to the combined effects of many genetic and environmental factors.*"

More than two variables

- Let's say that each event measure three quantities A, B and C
- We than have three random variables x, y and z
- Vector of measurements is now a matrix:

Event	A	B	C
1	x_1	y_1	z_1
2	x_2	y_2	z_2
...	...	...	...
N	x_N	y_N	z_N
Mean→	μ_x	μ_y	μ_z

- Introducing new notation

$$(x, y, z) \rightarrow (x_{(1)}, x_{(2)}, x_{(3)}) = \vec{x} = \boldsymbol{x}$$

$$(\mu_x, \mu_y, \mu_z) \rightarrow (\mu_{(1)}, \mu_{(2)}, \mu_{(3)}) = \vec{\mu} = \boldsymbol{\mu}$$

- In case of m variables $\boldsymbol{x} = (x_{(1)}, x_{(2)}, \ldots, x_{(m)})$
- Please note: this multivariate vector $\boldsymbol{x}$ is a vector of m variables for one event, while in the case of one variable $\boldsymbol{x}$ is a vector of values of one variable for N events

Multivariate Gaussian

- Multivariate Gaussian for the vector $\boldsymbol{x} = (x_{(1)}, x_{(2)}, \ldots, x_{(m)})$

$$f(\boldsymbol{x}; \boldsymbol{\mu}, V) = \frac{1}{(2\pi)^{n/2} |V|^{1/2}} \exp\left[-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^T V^{-1} (\boldsymbol{x}-\boldsymbol{\mu})\right]$$

- $\boldsymbol{x}$ and $\boldsymbol{\mu}$ are column vectors, while $\boldsymbol{x}^T$ and $\boldsymbol{\mu}^T$ are row vectors

$$\mu_{(i)} = E(x_{(i)}) \qquad V_{ij} = \mathrm{cov}[x_{(i)}, x_{(j)}]$$

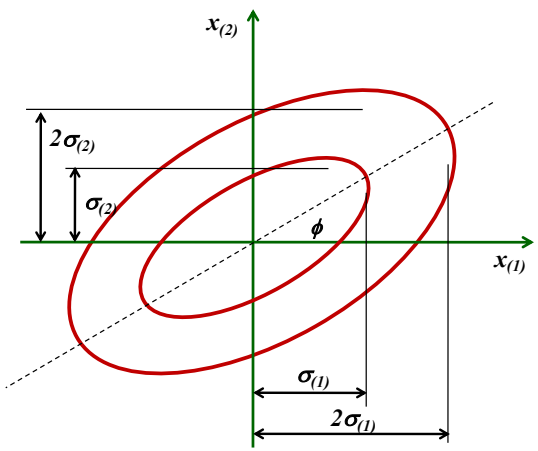
- Case of two variables ($m = 2$)

$$f(x_{(1)}, x_{(2)}; \mu_{(1)}, \mu_{(2)}, \sigma_{(1)}, \sigma_{(2)}) =$$

$$\frac{1}{2\pi\sigma_{(1)}\sigma_{(2)}\sqrt{1-\rho^2}} \times \exp\left\{-\frac{1}{2}\left[x_{(1)}-\mu_{(1)} \quad x_{(2)}-\mu_{(2)}\right]\begin{bmatrix}\sigma_{(1)}^2 & \rho\sigma_{(1)}\sigma_{(2)} \\ \rho\sigma_{(1)}\sigma_{(2)} & \sigma_{(2)}^2\end{bmatrix}^{-1}\begin{bmatrix}x_{(1)}-\mu_{(1)} \\ x_{(2)}-\mu_{(2)}\end{bmatrix}\right\} =$$

$$\frac{1}{2\pi\sigma_{(1)}\sigma_{(2)}\sqrt{1-\rho^2}} \times \exp\left\{-\frac{1}{2(1-\rho^2)}\left[\left(\frac{x_{(1)}-\mu_{(1)}}{\sigma_{(1)}}\right)^2 + \left(\frac{x_{(2)}-\mu_{(2)}}{\sigma_{(2)}}\right)^2 - 2\rho\left(\frac{x_{(1)}-\mu_{(1)}}{\sigma_{(1)}}\right)\left(\frac{x_{(2)}-\mu_{(2)}}{\sigma_{(2)}}\right)\right]\right\}$$

2D Gaussian: iso-probability curves



	P_{1D}	P_{2D}
1σ	0.6827	0.3934
2σ	0.9545	0.8647
3σ	0.9973	0.9889
1.515σ		0.6827
2.486σ		0.9545
3.439σ		0.9973

Remember (roughly) this values, we'll use them later in errors estimates!

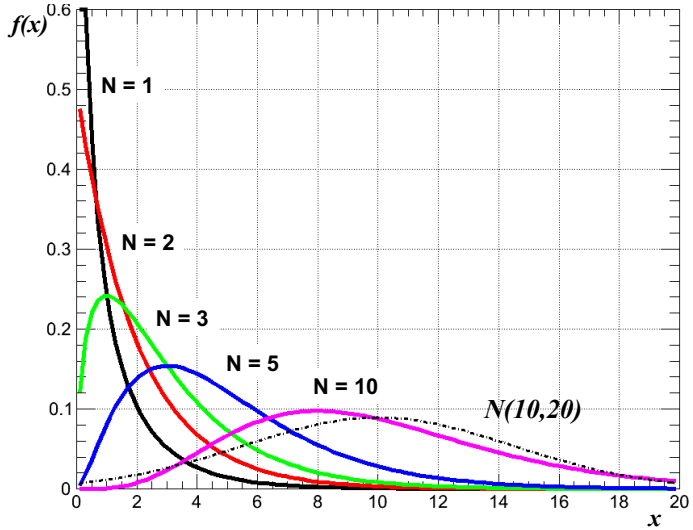
ϕ is a measure of the correlation (more details later)

Adopted from L. Lista

Chi-square distribution

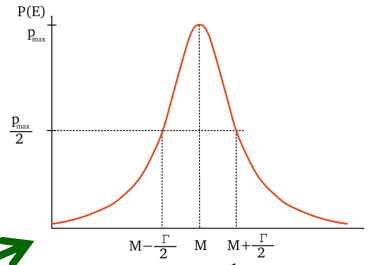
Variable	x, positive real number
Parameters	N, positive integer (number of "degrees of freedom")
Probability function	$f(x) = \left(\frac{1}{2}\left(\frac{X}{2}\right)^{\frac{N}{2}-1} e^{-\frac{x}{2}}\right) \Big/ \Gamma\left(\frac{N}{2}\right)$
Mean	$E(x) = N$
Variance	$V(x) = 2N$
Usage example	Chi-square test for goodness of fit
Comments	• If x_i are k independent, normally distributed random variables with mean 0 and variance 1, then the random variable $Q = \Sigma x_i^2$ is distributed according to the chi-square distribution with k degrees of freedom • The chi-square distribution is a special case of the gamma distribution.

Chi-square distribution



Some other distributions

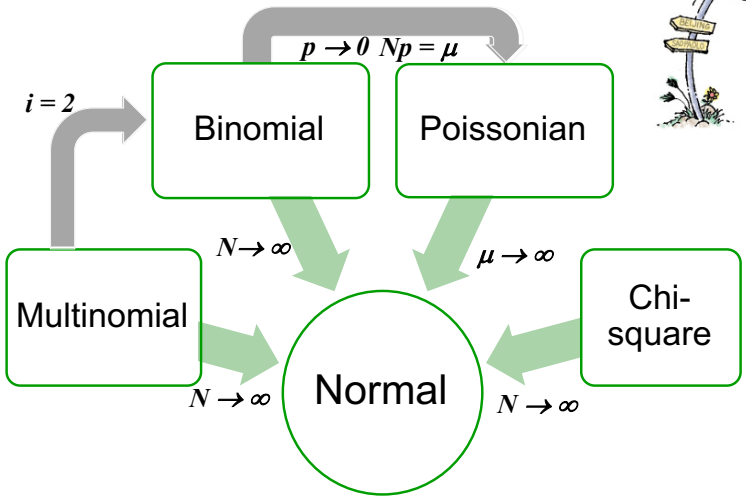
- **Student's t-distribution**
  - Used for hypothesis testing
  - First published in 1908 by W. S. Gosset, while he worked at a Guinness Brewery, under the pseudonym *Student*
- **Beta distribution**
  - Used in Bayesian statistics
- **Gamma distribution**
  - Probability model for waiting time
- **Cauchy or Lorentz or Breit-Wigner distribution**
  - A solution to the differential equation describing a **resonance**
  - Energy distribution of a resonance



$$P(E) \sim \frac{1}{(E^2 - M^2)^2 + M^2\Gamma^2}$$

- **Log-Normal distribution**
  - Used when including systematic errors in the analysis
  - If x is Log-Normally distributed, than $log(x)$ is Normally distributed

All roads lead to Rome



General picture

1 **Physical phenomena** *Described by a theory*

2 Sampling a reality **Experiment**

3 **Data sample** $\boldsymbol{x} = (x_1, x_2, \ldots, x_N)$

For example:
$\boldsymbol{x} = (event_1, \ldots, event_N)$

4 Data **analysis**

5 **Results**

- parameter estimates
- confidence limits
- hypothesis tests

Described by PDFs, depending on p **uknown** parameters with **true values**
$\boldsymbol{\theta}^{true} = (\theta_1^{true}, \theta_2^{true}, \ldots, \theta_p^{true})$
For example:
$\boldsymbol{\theta}^{true} = (m_H^{true}, \Delta m_s^{true}, \ldots, \sigma_{tot}^{true})$

In statistics $\boldsymbol{x}$ is a multivariate **random variable** *(each event has many properties, all potential variables)*

Physicists and statisticians

- Example: histogram fitting

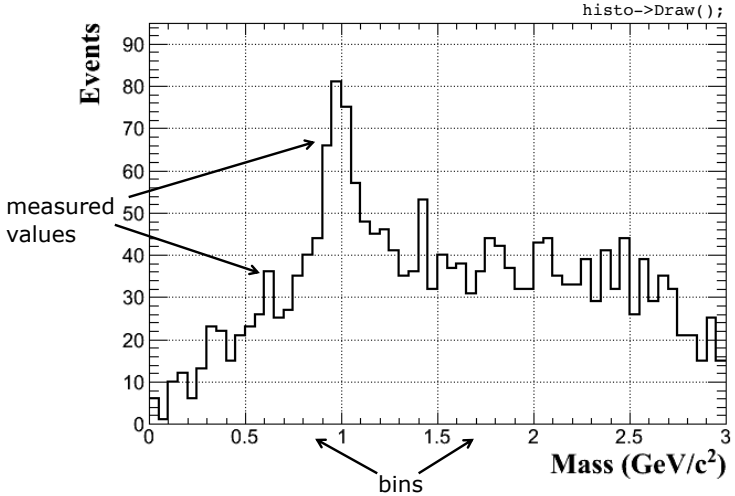
Physicists	Statisticians
1. Determining the "best fit" parameters of a curve	1. Point estimation
2. Determining the errors on the parameters	2. Confidence interval estimation
3. Judging the goodness of a fit	3. Goodness-of-fit (Hypothesis) testing

Adopted from [Baker, Cousins, 1984]

From $ROOTSYS/tutorials/fit/FittingDemo.C

Example: mass measurement

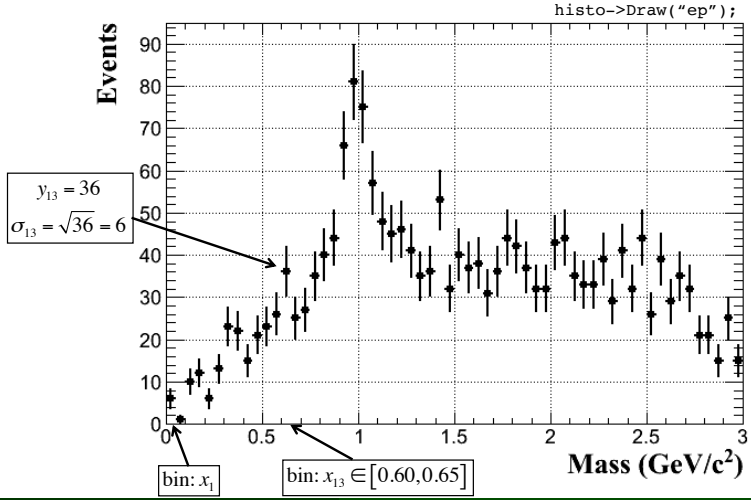
Histogram: default representation



From $ROOTSYS/tutorials/fit/FittingDemo.C

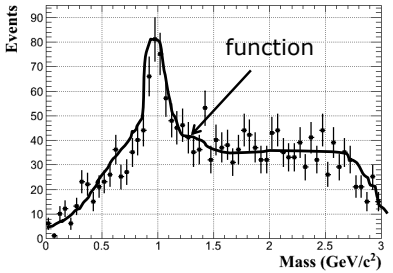
Example: mass measurement

Histogram: points with errors



Example: mass measurements

- Therefore we have
  - a set of precisely known values $\boldsymbol{x} = (x_1,\ldots,x_N)$ - *histograms bins*
  - At each $\boldsymbol{x}_i$
    - a measured value $\boldsymbol{y}_i$ - *number of events in a given bin*
    - a corresponding error on measured value σ_i



1 **Physical phenomena** ***(theory)***

Described by a function, depending on p **uknown** parameters with **true values**

$$\boldsymbol{\theta}^{true} = (\theta_1^{true}, \theta_2^{true}, \ldots, \theta_p^{true})$$

- What we want is to estimate the values of θ_i^{true}
- This is what we call **the parameter ESTIMATOR**: $\hat{\theta}_i$

Properties of a good estimator

- **Consistent**
  - Estimate coverges to the true value as amount of data increases

$$\hat{\theta} \xrightarrow{N \text{ increases}} \theta^{true}$$

- **Unbiased**
  - Bias is the difference between expected value of the estimator and the true value of the parameter

$$b = E(\hat{\theta}) - \theta^{true} = 0$$

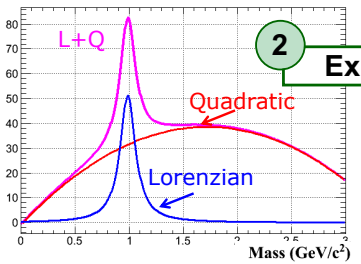
- **Efficient**
  - Cramér-Rao bound for the minimum of the variance of estimator:
  - Estimator is efficient when its variance reaches the lower bound

$$V(\hat{\theta}) = \frac{\left(1 + \frac{\partial b}{\partial \theta}\right)^2}{\underbrace{E\left[\left(\frac{\partial}{\partial \theta}\sum_i \ln f(x_i;\theta)\right)\right]}_{\text{Fisher information}}}$$

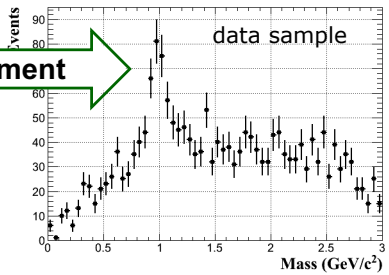
- **Robust**
  - Insensitive to departures from assumptions in the PDF

Experiment: sampling the reality

- Physical phenomena



2 **Experiment**

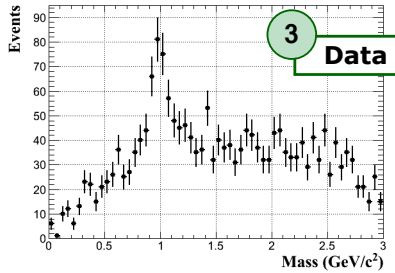


$$Lorenzian = L(x;D,\Gamma,M) \sim \frac{D\Gamma}{(x^2 - M^2)^2 + 0.25\Gamma^2} \qquad Quadratic = Q(x;A,B,C) \sim A + Bx + Cx^2$$

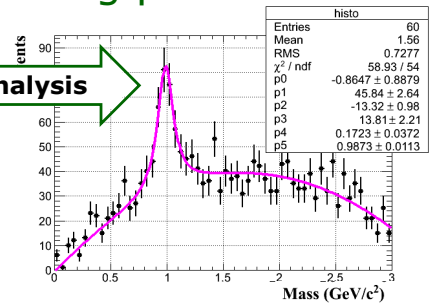
- Underlying phenomena depends on 6 unknown parameters:

$$F(x;D,\Gamma,M,A,B,C) = L(x;D,\Gamma,M) + Q(x;A,B,C) = F(x;\theta)$$

Data analysis: estimating parameters



3 **Data analysis**



- From data sample we are looking for the **function** that describes the measurements the best
- The parameters of that **function** are estimators of uknown parameters

$$F(x;\hat{D},\hat{\Gamma},\hat{M},\hat{A},\hat{B},\hat{C}) = L(x;\hat{D},\hat{\Gamma},\hat{M}) + Q(x;\hat{A},\hat{B},\hat{C}) = F(x;\hat{\theta})$$

Statistic

- Be carefull: **statistic** is not statistic**S**!
- Any new random variable (f.g. T), defined as a function of a measured sample x is called a **statistic**

$$T = T(x_1, \ldots, x_N)$$

- For example, the sample mean $\bar{x} = \frac{1}{N}\sum_{i=1}^{N} x_i$ is a statistic!
- A statistic used to estimate a parameter is called an **estimator**
  - For instance, the ***sample mean*** is a statistic and an estimator for the ***population mean***, which is an uknown parameter
  - **Estimator** is a function of the data
  - **Estimate**, a value of estimator, is our "best" guess for the true value of parameter
- Some other example of statistics: sample median, variance, standard deviation, quartiles, percentiles, t-statistic, chi-square statistic, kurtosis, skewness etc.

Estimators in ROOT - values

Mean	RMS (it's actually σ, name RMS is historic)
$\frac{1}{N}\sum_{i=1}^{N} x_i \pm \frac{RMS}{\sqrt{N}}$	$\frac{1}{N}\sum_{i=1}^{N}(x_i-\mu)^2 \pm \frac{RMS}{\sqrt{2N}}$
Skewness	**Kurtosis**
$\frac{1}{N}\sum_{i=1}^{N}(x_i-\bar{x})^3 \Big/ \left(\frac{1}{N}\sum_{i=1}^{N}(x_i-\bar{x})^2\right)^{3/2} \pm \sqrt{\frac{6}{N}}$	$\frac{1}{N}\sum_{i=1}^{N}(x_i-\bar{x})^4 \Big/ \left(\frac{1}{N}\sum_{i=1}^{N}(x_i-\bar{x})^2\right)^{4} - 3 \pm \sqrt{\frac{24}{N}}$

- Total number of events N is only in the currently defined range
- From the ROOT Reference Manual

  "Note that the mean value/RMS is computed using the bins in the currently defined range (see TAxis::SetRange). By default the range includes all bins from 1 to nbins included, excluding underflows and overflows. To force the underflows and overflows in the computation, one must call the static function TH1::StatOverflows(kTRUE) before filling the histogram."

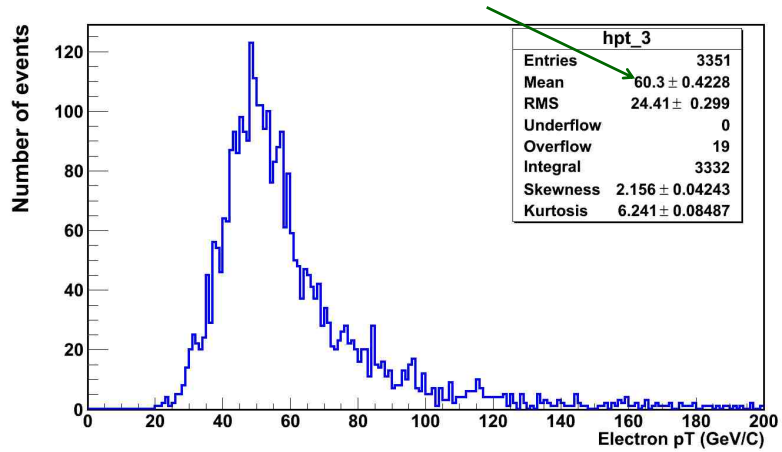
Estimators in ROOT - display

- Estimators display in the statistic box
  - Drawn by default; can be eleminated by `TH1::SetStats(kFALSE)`
- `gStyle->SetOptStat(mode)` allows to select the type of displayed information
  - `mode = ksiourmen (default = 000001111)`

n = 1	the name of histogram is printed
e = 1	the number of entries
m = 1	the mean value
m = 2	the mean and mean error values
r = 1	the root mean square (RMS)
r = 2	the RMS and RMS error
u = 1	the number of underflows
o = 1	the number of overflows
i = 1	the integral of bins
s = 1	the skewness
s = 2	the skewness and the skewness error
k = 1	the kurtosis
k = 2	the kurtosis and the kurtosis error

Estimators in ROOT - example

Notice an influence of the tail on the mean value



How to find a good estimator?

The Method of Moments

- Giving consistent and asymptotically unbiased estimators
- But are not as efficient as the maximum likelihood estimates
- Not covered in this lecture

The Maximum Likelihood Method

- Also giving consistent and asymptotically unbiased estimators
- Widely used in practice

The Least Squares Method (Chi-Square)

- Giving consistent estimator
- Linear chi-square estimator is unbiased
- Frequently used in histogram fitting

Likelihood function

- Assume that observations (events) are independent
  - With the PDF depending on parameters $\boldsymbol{\theta}: f(x_i;\boldsymbol{\theta})$
- **The probability that all N events will happen**, i.e. the PDF of $\boldsymbol{x}$ is, by independence, a product of all single events PDFs

$$P(\boldsymbol{x};\boldsymbol{\theta}) = P(x_1,\ldots,x_N;\boldsymbol{\theta}) = \prod_{i=1}^{N} f(x_i;\boldsymbol{\theta})$$

- When the variable $\boldsymbol{x}$ is replaced by the observed data $\boldsymbol{x}^0$, then P is no longer a PDF
- It is usual to denote it by L and call $L(\boldsymbol{X}^0;\boldsymbol{\theta})$ the **<u>likelihood function</u>**
  - Which is now a function of $\boldsymbol{\theta}$ only $\quad L(\boldsymbol{\theta}) = P(\boldsymbol{X}^0;\boldsymbol{\theta})$
- Often in the literature, and through this lectures, it's convenient to keep $\boldsymbol{X}$ as a variable and continue to use notation $L(\boldsymbol{X};\boldsymbol{\theta})$

Maximum likelihood method

- Reminder: the probability that all N independent events will happen is given by the **likelihood function** $L(\boldsymbol{x};\boldsymbol{\theta}) = \prod_{i=1}^{N} f(x_i;\boldsymbol{\theta})$
- The principle of maximum likelihood (ML) says:

The maximum likelihood estimator $\hat{\theta}$ is the value of θ for which the likelihood is a maximum!

- In words of R. J. Barlow: *"You determine the value of θ that makes the probability of the actual results obtained, $\{x_1, \ldots, x_N\}$, as large as it can possible be."*
- In practice it's easier to maximize the **log-likelihood function**

$$\ln L(\boldsymbol{x};\boldsymbol{\theta}) = \sum_{i=1}^{n} \ln f(x_i;\boldsymbol{\theta})$$

- For p parameters we get a set of p likelihood equations

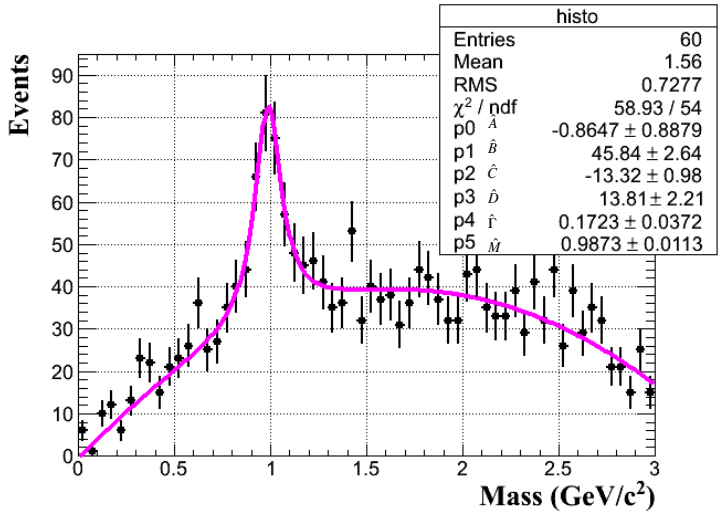
$$\frac{\partial \ln L(\boldsymbol{x};\boldsymbol{\theta})}{\partial \theta_j} = 0, \quad j = 1,2,\ldots,p$$

- It is often more convenient the **minimize $-\ln L$** or **$-2\ln L$**
  - Minimization with MINUIT/MIGRAD or FUMILI in ROOT

Maximum Likelihood - comments

- ML estimator is **consistent**
- ML estimate is approximately **unbiased** and **efficient** for large samples
  - Still usefull for small samples, but with extra care!
- ML estimate is **invariant**
  - A transformation of parameter won't change the answer
- **ML estimate is not the most likely value of parameter; it is the estimate that makes your data the most likely!**
- What was presented up to now is sometimes called **unbinned maximum likelihood**
- **Binned maximum likelihood**: when data are organized in bins
  - See "ML fit of a histogram" later on
- Extra care to be taken when the best value of parameters are near imposed limits
- ML has many advantages, but a few drawbacks too
  - F.g. goodness-of-fit for ML is non-trivial issue, still open and debated

Example: results of the fit



ADVANCED

Example – ML fit of a histogram (1/2)

- Suppose one has
  - N events in a histogram with k bins
  - n_i in the ith bin → vector of data $\boldsymbol{n} = (n_1, \ldots, n_k)$
  - Expected number of events in each bin depend on uknown parameters $\boldsymbol{\theta}$, $\boldsymbol{\nu}(\boldsymbol{\theta}) = (\nu_1, \ldots, \nu_k)$
  - Given ν_i probability to have n_i is $f(n_i; \nu_i)$
    - Usually probability is Poissonian:

$$f(n_i;\nu_i) = \frac{\nu_i^{n_i} e^{-\nu_i}}{n_i!}$$

- The likelihood function is $L(\boldsymbol{n};\boldsymbol{\nu}) = \prod_i \frac{\nu_i^{n_i} e^{-\nu_i}}{n_i!}$
- To find best estimate of θ we have to maximize $\ln L(\boldsymbol{n};\boldsymbol{\nu})$ based on the contents of the bins

ADVANCED

Example – ML fit of a histogram (2/2)

- In can be shown that this procedure is equivalent to maximizing the **likelihood ratio**

$$\lambda(\boldsymbol{\theta}) = \frac{L(\boldsymbol{n};\boldsymbol{\nu}(\boldsymbol{\theta}))}{L(\boldsymbol{n};\boldsymbol{m})} \approx \frac{L(\boldsymbol{n};\boldsymbol{\nu}(\boldsymbol{\theta}))}{L(\boldsymbol{n};\boldsymbol{n})}$$

  - Where $\boldsymbol{m} = (m_1, \ldots, m_k)$ are true (uknown) values of $\boldsymbol{n}$
  - Best bin-to-bin model independent maximum likelihood estimate of $\boldsymbol{m}$ is actually $\boldsymbol{n}$
- Maximizing $\lambda(\boldsymbol{\theta})$ is equivalent to **minimizing**

$$-2\ln\lambda(\boldsymbol{\theta}) = 2\sum_{i=1}^{N}\left[\nu_i(\boldsymbol{\theta}) - n_i + n_i \ln\frac{n_i}{\nu_i(\boldsymbol{\theta})}\right]$$

  - Which is now much easier to implement then maximizing $\ln L(\boldsymbol{n};\boldsymbol{\nu})$
- In case where $n_i = 0$, last term in eq. above is zero

ADVANCED

Extended maximum likelihood

- In the usual maximum likelihood method
  - Parameter relevant to the **shapes** of distributions are determined
  - Absolute **normalization** is **equal** to the **observed** number of events
- If we want to **estimate** the **absolute normalization** the so called **"Extended maximum likelihood method"** is used
- Example: From the vector of measurements $\boldsymbol{x} = (x_1, \ldots, x_N)$ we want to estimate number of signal events (s), number of background events (b) and a vector of parameters $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_p)$
- Likelihood function is

$$L(\boldsymbol{x};s,b,\boldsymbol{\theta}) = \frac{(s+b)^N e^{-(s+b)}}{N!}\prod_{i=1}^{N}\left(\frac{s}{s+b}P_s(x_i;\boldsymbol{\theta}) + \frac{b}{s+b}P_b(x_i;\boldsymbol{\theta})\right)$$

- To obtain s, b and $\boldsymbol{\theta}$ we maximize (or mimimize $-2\ln L$)

Constant

$$\ln L(\boldsymbol{x};s,b,\boldsymbol{\theta}) = -s - b + \sum_{i=1}^{N}\ln\left(\frac{s}{s+b}P_s(x_i;\boldsymbol{\theta}) + \frac{b}{s+b}P_b(x_i;\boldsymbol{\theta})\right) - \ln(N!)$$

Least squares method

- Suppose we have
  - A set of precisely known values $\boldsymbol{x} = (x_1,\ldots,x_N)$
    - *For example histograms bins*
  - At each x_i
    - a measured value y_i
      - *For example number of events in the given histogram bin*
    - corresponding error on measured value σ_i
    - predicted value of measurement that depends on parameters $\boldsymbol{\theta} = (\theta_1,\ldots,\theta_p)$ we want to estimate: $F(x_i;\boldsymbol{\theta})$
  - Suppose that measurements are independent
- To find best estimate of θ we minimize the suitably weighted summ of squared differences between measured and predicted values → so called "**least squares**" or "**chi-square**"

$$\chi^2(\boldsymbol{\theta}) = \sum_{i=1}^{N} \frac{\left(y_i - F(x_i;\boldsymbol{\theta})\right)^2}{\sigma_i^2}$$

Least squares method

- If y_i are Gaussian distributed with variances σ_i

$$\chi^2(\boldsymbol{\theta}) = \sum_{i=1}^{N} \frac{\left(y_i - F(x_i;\boldsymbol{\theta})\right)^2}{\sigma_i^2} = -2\ln L(\boldsymbol{\theta}) + \text{constant}$$

Minimizing chi-square χ^2

↕

Maximizing log-likelihood $\ln L$

or minimizing $-2\ln L$

Pearson's vs Neyman's chi-square

- If y_i are Poissonian distributed, there are two choices
  - Reminder first: for Poissonian **variance = mean value** ($\sigma^2 = \mu$)
  - So called **Pearson's chi-square** (or "**chi-square**")

$$\chi^2(\boldsymbol{\theta}) = \sum_{i=1}^{N} \frac{\left(y_i - F(x_i;\boldsymbol{\theta})\right)^2}{F(x_i;\boldsymbol{\theta})}$$

    - But now σ_i depends on $\boldsymbol{\theta}$ which complicates the minimization
  - So called **Neyman's chi-square** (or "**modified chi-square**")

$$\chi^2(\boldsymbol{\theta}) = \sum_{i=1}^{N} \frac{\left(y_i - F(x_i;\boldsymbol{\theta})\right)^2}{y_i}$$

  - Minimization simpler
  - Easier to combine data with different basic accuracies
  - Problem with $y_i = 0$
    - For example in ROOT this bin ignored
    - For small samples better use ML
- The best values of parameters $\theta = (\theta_1,\ldots,\theta_p)$ are found by solving p equations

$$\frac{\partial \chi^2(\boldsymbol{\theta})}{\partial \theta_i} = 0, \quad i = 1,\ldots,p$$