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**
    \[
    V(x) = \sigma^2 = E[(x - \mu)^2] = E(x^2) - \mu^2 = \int (x - \mu)^2 f(x) dx
    \]
  - \( \sigma \) is called the standard deviation
- \( E(x) \) is a measure of the location of the distribution
- \( V(x) \) is a measure of the spread of the distribution

Moments
- \( \mu_n = E(x^n) \) is the \( n^{th} \) algebraic moment
- \( V_n = E\{[x^n - E(x)]^n\} \) is the \( n^{th} \) central moment
- \( \mu'_n = E\{|x^n|\} \) is the \( n^{th} \) absolute moment
- \( V'_n = E\{|x^n - E(x)|^n\} \) is the \( n^{th} \) absolute central moment

**The coefficient of skewness**
A measure of the skewness of the distribution
\[
\gamma_1 = \frac{V_3}{V_2^{3/2}}
\]

**The coefficient of kurtosis**
A measure of the "peakedness" of the distribution
\[
\gamma_2 = \frac{V_4}{V_2^2} - 3
\]
Covariances and correlations

- Joint PDF for two random variables is\( f(x,y) \).
- The **mean** and the **variance** of \( x \) and \( y \):
  \[
  \mu_x = E(x) = \int xf(x,y)dx \quad \mu_y = E(y) = \int yf(x,y)dx \\
  \sigma_x^2 = E[(x - \mu_x)^2] \quad \sigma_y^2 = E[(y - \mu_y)^2] 
  \]
- **Covariance** \( \text{cov}(x,y) = E[(x - \mu_x)(y - \mu_y)] = E(xy) - E(x)E(y) \)
- **Correlation coefficient** \( \text{corr}(x,y) = \rho(x,y) = \frac{\text{cov}(x,y)}{\sigma_x \sigma_y} \)

**Covariance/Variance/Error matrix:**
\[
V = \begin{bmatrix}
\text{cov}(x,x) & \text{cov}(x,y) \\
\text{cov}(x,y) & \text{cov}(y,y)
\end{bmatrix} =
\begin{bmatrix}
\sigma_x^2 & \rho_{xy} \sigma_x \sigma_y \\
\rho_{xy} \sigma_x \sigma_y & \sigma_y^2
\end{bmatrix}
\]

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** - \( Z \) 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 \( r \) successes in \( N \) trials, when probability of success in each single trial is a constant, \( p \)

Multinomial distribution

- **Variable** \( r_i \), \( i = 1, \ldots, k \), positive integers \( \leq N \)
- **Parameters** \( N \), positive integer; \( k \), positive integer \( p_i \geq 0, \ i = 1, \ldots, k \), \( \sum p_i = 1 \)
- **Probability function**
  \[ P(r_1,\ldots,r_k;N,p_1,\ldots,p_k) = \frac{N!}{r_1!\cdots r_k!} p_1^{r_1} \cdots p_k^{r_k} \]
- **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 \( i^{th} \) 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**  
\( \mu \), 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 \( \int L dt \). Expected number of events is \( \mu = \sigma \int L dt \). \( \sigma \) 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 independently of the time since the last event.
- \( \mu \) represents expected number of events in a given time interval.
- Time between two successive events is exponentially distributed.
- Poisson distribution is also called Poissonian.

### Normal or Gaussian distribution

**Variable**  
\( x \), positive real number

**Parameters**  
\( \mu \), real number  
\( \sigma \), real number

**Probability density function**  
\[
f(x) = N(\mu, \sigma^2) = \frac{1}{\sqrt{2\pi} \sigma} e^{-\frac{(x - \mu)^2}{2 \sigma^2}}
\]

**Mean**  
\( E(x) = \mu \)

**Variance**  
\( V(x) = \sigma^2 \)

**Cumulative distribution**  
\[
F(x) = \Phi \left( \frac{x - \mu}{\sigma} \right), \quad \Phi(z) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{z} e^{-\frac{t^2}{2}} dt
\]

**Comments**
- The most important distribution in statistics.
- The half-width at half-height is 1.176\( \sigma \).
- \( N(0, 1) \) is called standard Normal density.
- Any linear combination of the \( x_i \) is also Normal.

### Gaussian – some properties

<table>
<thead>
<tr>
<th>( n )</th>
<th>Area ( \pm 1\sigma )</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.682689492137</td>
</tr>
<tr>
<td>2</td>
<td>0.954499736104</td>
</tr>
<tr>
<td>3</td>
<td>0.997300203937</td>
</tr>
<tr>
<td>4</td>
<td>0.999936657516</td>
</tr>
<tr>
<td>5</td>
<td>0.999999426697</td>
</tr>
</tbody>
</table>
Why is Gauss Normal?

Central limit theorem:
If we have a set of N independent variables $x_i$, each from a distribution with mean $\mu_i$ and variance $\sigma_i^2$, then the distribution of the sum $X = \sum x_i$
- 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 (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."

Multivariate Gaussian

Multivariate Gaussian for the vector $x = (x_1, x_2, \ldots, x_m)$

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

$x$ and $\mu$ are column vectors, while $x^T$ and $\mu^T$ are row vectors.

$\mu = E(x_i)$

$V = \text{cov}(x_i, x_j)$

Case of two variables ($m = 2$)

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

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

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:

Introducing new notation

$$(x_1, y_1, z_1) \rightarrow (x_{(1)}, x_{(2)}, x_{(3)}) \rightarrow \bar{x} = x$$

$$(\mu_1, \mu_2, \mu_3) \rightarrow (\mu_{(1)}, \mu_{(2)}, \mu_{(3)}) \rightarrow \bar{\mu} = \mu$$

In case of $m$ variables $x = (x_{(1)}, x_{(2)}, \ldots, x_{(m)})$

Please note: this multivariate vector $x$ is a vector of $m$ variables for one event, while in the case of one variable $x$ is a vector of values of one variable for $N$ events

2D Gaussian: iso-probability curves

$\phi$ is a measure of the correlation (more details later)

Adopted from L. Lista
### Chi-square distribution

<table>
<thead>
<tr>
<th>Variable</th>
<th>$x$, positive real number</th>
</tr>
</thead>
<tbody>
<tr>
<td>Parameters</td>
<td>$N$, positive integer (number of &quot;degrees of freedom&quot;)</td>
</tr>
<tr>
<td>Probability function</td>
<td>$f(x) = \left( \frac{X^2}{2} \right)^{N/2} e^{-X/2} / \Gamma(N/2)$</td>
</tr>
<tr>
<td>Mean</td>
<td>$E(x) = N$</td>
</tr>
<tr>
<td>Variance</td>
<td>$V(x) = 2N$</td>
</tr>
<tr>
<td>Usage example</td>
<td>Chi-square test for goodness of fit</td>
</tr>
</tbody>
</table>
| Comments       | • If $x_i$ are $k$ independent, normally distributed random variables with mean 0 and variance 1, then the random variable $Q = \sum x_i^2$ 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. |

### 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

- **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

- **Binomial**
- **Poissonian**

- **Multinomial**

- **Normal**

- **Chi-square**

- $$p \to 0 \quad Np = \mu$$

- $$N \to \infty$$

- $$\mu \to \infty$$
General picture

1. Physical phenomena
   Described by a theory

   \[ W_{\text{MC}} - W_{\text{MC}} \]

2. Sampling a reality
   Experiment

3. Data sample
   \[ x = (x_1, x_2, \ldots, x_n) \]
   For example:
   \[ x = (\text{event}_1, \ldots, \text{event}_n) \]

4. Data analysis

5. Results
   - parameter estimates
   - confidence limits
   - hypothesis tests

Data analysis

- **Physical phenomena**
  - Described by PDFs, depending on \( p \) unknown parameters with true values
  - \( \theta^{p\text{true}} = (\theta_1^{p\text{true}}, \theta_2^{p\text{true}}, \ldots, \theta_p^{p\text{true}}) \)
  - For example:
    - \( \theta^{p\text{true}} = (m_i^{p\text{true}}, \Delta m_i^{p\text{true}}, \sigma_i^{p\text{true}}) \)

- **Experiment**
  - Data sample
    \[ x = (x_1, x_2, \ldots, x_n) \]

- **Results**
  - Parameter estimates
  - Confidence limits
  - Hypothesis tests

Physicists and statisticians

**Example: histogram fitting**

<table>
<thead>
<tr>
<th><strong>Physicists</strong></th>
<th><strong>Statisticians</strong></th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Determining the “best fit” parameters of a curve</td>
<td>1. Point estimation</td>
</tr>
<tr>
<td>2. Determining the errors on the parameters</td>
<td>2. Confidence interval estimation</td>
</tr>
<tr>
<td>3. Judging the goodness of a fit</td>
<td>3. Goodness-of-fit (Hypothesis) testing</td>
</tr>
</tbody>
</table>

Adapted from [Baker, Cousins, 1984]

Example: mass measurement

**Example: mass measurement**

Histogram: default representation

```
histo->Draw();
```

Histogram: points with errors

```
histo->Draw("ep");
```

From `$ROOTSYS/tutorials/fit/FittingDemo.C`

From `$ROOTSYS/tutorials/fit/FittingDemo.C`
Example: mass measurements

Therefore we have
- a set of precisely known values \( x = (x_1, \ldots, x_n) \) - histograms bins
- At each \( x_j \)
  - a measured value \( y_j \) - number of events in a given bin
  - a corresponding error on measured value \( \sigma_i \)

What we want is to estimate the values of \( \theta_1^{\text{true}}, \ldots, \theta_m^{\text{true}} \)

This is what we call the parameter ESTIMATOR: \( \hat{\theta}_j \)

Physical phenomena (theory)

Described by a function, depending on \( m \) unknown parameters with true values \( \theta_1^{\text{true}}, \ldots, \theta_m^{\text{true}} \)

Properties of a good estimator

Consistent
- Estimate converges to the true value as amount of data increases
  \( \hat{\theta} \) increases \( \Rightarrow \theta^{\text{true}} \)

Unbiased
- Bias is the difference between expected value of the estimator and the true value of the parameter
  \( b = E(\hat{\theta}) - \theta^{\text{true}} = 0 \)

Efficient
- Cramér-Rao bound for the minimum of the variance of estimator:
  \[
  V(\hat{\theta}) = \frac{1}{E \left( \frac{\partial^2}{\partial \theta^2} \sum \ln f(x; \theta) \right)}
  \]

Robust
- Insensitive to departures from assumptions in the PDF
- Fisher information

Experiment: sampling the reality

Physical phenomena

L+Q
Lorenzian
Quadratic

\[
\text{Lorenzian} = L(x; D, \Gamma, M) = \frac{D \Gamma}{(x^2 - M^2)^2 + 0.55D^2}
\]

\[
\text{Quadratic} = Q(x; A, B, C) = A + Bx + Cx^2
\]

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

From data sample we are looking for the function that describes the measurements the best

The parameters of that function are estimators of unknown parameters

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

- Be careful! **Statistic** is not statistic!
- 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 unknown 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 - display

- **Estimators** display in the statistic box
  - Drawn by default; can be eliminated by \( \text{TH1::SetStats(kFALSE)} \)
  - \( \text{gStyle} \rightarrow \text{SetOptStat(mode)} \) allows to select the type of displayed information
    - mode: 
      - \( kTRUE \) (default - 00000111)

<table>
<thead>
<tr>
<th>n</th>
<th>the name of histogram is printed</th>
</tr>
</thead>
<tbody>
<tr>
<td>e</td>
<td>the number of entries</td>
</tr>
<tr>
<td>m</td>
<td>the mean value</td>
</tr>
<tr>
<td>m</td>
<td>the mean and mean error values</td>
</tr>
<tr>
<td>r</td>
<td>the root mean square (RMS)</td>
</tr>
<tr>
<td>r</td>
<td>the RMS and RMS error</td>
</tr>
<tr>
<td>u</td>
<td>the number of underflows</td>
</tr>
<tr>
<td>o</td>
<td>the number of overflows</td>
</tr>
<tr>
<td>i</td>
<td>the integral of bins</td>
</tr>
<tr>
<td>s</td>
<td>the skewness</td>
</tr>
<tr>
<td>s</td>
<td>the skewness and the skewness error</td>
</tr>
<tr>
<td>k</td>
<td>the kurtosis</td>
</tr>
<tr>
<td>k</td>
<td>the kurtosis and the kurtosis error</td>
</tr>
</tbody>
</table>

Estimators in ROOT - values

<table>
<thead>
<tr>
<th>Mean</th>
<th>RMS</th>
</tr>
</thead>
<tbody>
<tr>
<td>( \frac{1}{N} \sum_{i=1}^{N} x_i \pm \frac{\text{RMS}}{\sqrt{N}} )</td>
<td>( \frac{1}{N} \sum_{i=1}^{N} (x_i - \mu)^2 \pm \frac{\text{RMS}}{\sqrt{2N}} )</td>
</tr>
</tbody>
</table>

- **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 \( \text{TAxis::SetRange} \)). 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 \( \text{TH1::StatOverflows(kTRUE)} \) before filling the histogram."

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 \( \theta : f(x; \theta) \)

- **The probability that all \( N \) events will happen**, i.e. the PDF of \( x \) is, by independence, a product of all single events PDFs
  \[
  P(x; \theta) = P(x_1, \ldots, x_N; \theta) = \prod_{i=1}^{N} f(x_i; \theta)
  \]

- When the variable \( x \) is replaced by the observed data \( x^0 \), then \( P \) is no longer a PDF

- It is usual to denote it by \( L \) and call \( L(x^0; \theta) \) the **likelihood function**
  - Which is now a function of \( \theta \) only
  \[
  L(\theta) = P(x^0; \theta)
  \]

- Often in the literature, and throughout this lectures, it’s convenient to keep \( X \) as a variable and continue to use notation \( L(X; \theta) \)

---

Maximum likelihood method

- Reminder: the probability that all \( N \) independent events will happen is given by the **likelihood function**
  \[
  L(x; \theta) = \prod_{i=1}^{N} f(x_i; \theta)
  \]

- The principle of maximum likelihood (ML) says:
  - **The maximum likelihood estimator** \( \hat{\theta} \) is the value of \( \theta \) for which the likelihood is a maximum!

- In words of R. J. Barlow: "You determine the value of \( \theta \) that makes the probability of the actual results obtained, \( \{x_1, \ldots, x_N\} \), as large as it can possibly be."

- In practice it’s easier to maximize the **log-likelihood function**
  \[
  \ln L(x; \theta) = \sum_{i=1}^{N} \ln f(x_i; \theta)
  \]

- For \( p \) parameters we get a set of \( p \) likelihood equations
  \[
  \frac{\partial \ln L(x; \theta)}{\partial \theta_j} = 0, \quad j = 1, 2, \ldots, p
  \]

- It is often more convenient the **minimize** \(-\ln L\) or \(-2\ln L\)
  - Minimization with MINUIT/MIGRAD or FUMILI in ROOT
Example: results of the fit

<table>
<thead>
<tr>
<th>Entries</th>
<th>60</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean</td>
<td>1.56</td>
</tr>
<tr>
<td>RMS</td>
<td>0.7277</td>
</tr>
</tbody>
</table>

\[ \chi^2 / \text{ndf} = 58.93 / 54 \]

p0 \[ \bar{a} = -0.8647 \pm 0.8879 \]
p1 \[ \bar{b} = 45.84 \pm 2.64 \]
p2 \[ \bar{c} = -13.32 \pm 0.98 \]
p3 \[ \bar{d} = 13.81 \pm 2.21 \]
p4 \[ \bar{e} = 0.1723 \pm 0.0372 \]
p5 \[ \bar{f} = 0.9873 \pm 0.0113 \]

---

Example – ML fit of a histogram (1/2)

Suppose one has:

- N events in a histogram with k bins
- \( n_i \) in the \( i \)th bin \( \rightarrow \) vector of data \( n = (n_1, ..., n_k) \)
- Expected number of events in each bin depend on unknown parameters \( \hat{\theta}, \hat{\nu} = (\nu_1, ..., \nu_q) \)
- Given \( \nu \), probability to have \( n_i \) is \( f(n_i; \nu) \)
  - Usually probability is Poissonian:
    \[ f(n_i; \nu) = \frac{\nu^{n_i} e^{-\nu}}{n_i!} \]

The likelihood function is:
\[ L(n; \nu) = \prod_i^n \frac{\nu^{n_i} e^{-\nu}}{n_i!} \]

To find best estimate of \( \hat{\theta} \) we have to maximize \( \ln L(n; \nu) \) based on the contents of the bins

---

Example – ML fit of a histogram (2/2)

In can be shown that this procedure is equivalent to maximizing the likelihood ratio
\[ \hat{\lambda}(\theta) = \frac{L(n; \nu(\theta))}{L(n; \nu)} = \frac{L(n; \nu(\theta))}{L(n; \nu(m))} \]

- Where \( m = (m_1, ..., m_k) \) are true (unknown) values of \( n \)
- Best bin-to-bin model independent maximum likelihood estimate of \( m \) is actually \( n \)

Maximizing \( \hat{\lambda}(\theta) \) is equivalent to minimizing
\[ -2 \ln \hat{\lambda}(\theta) = 2 \ln(L(n; \nu(\theta))) - n_i \ln \frac{n_i}{\nu_i(\theta)} \]

Which is now much easier to implement than maximizing \( \ln L(n; \nu) \)

In case where \( n_i = \theta \), last term in eq. above is zero

---

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 \( x = (x_1, ..., x_N) \) we want to estimate number of signal events \( (s) \), number of background events \( (b) \) and a vector of parameters \( \theta = (\theta_1, ..., \theta_q) \)

Likelihood function is
\[ L(x; s, b, \theta) = \frac{(s+b)^N e^{-(s+b)}}{N!} \prod_{i=1}^N \left( \frac{s}{s+b} P_i(x_i; \theta) + \frac{b}{s+b} P_i(x_i; \theta) \right) \]

To obtain \( s, b \) and \( \theta \) we maximize (or minimize \(-2 \ln L\))

\[ \ln L(x; s, b, \theta) = -s - b + \sum_{i=1}^N \ln \left( \frac{s}{s+b} P_i(x_i; \theta) + \frac{b}{s+b} P_i(x_i; \theta) \right) - \ln(N!) \]
Least squares method

Suppose we have
- A set of precisely known values \( x = (x_1, \ldots, x_N) \)
  - 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 \( s_i \)
- predicted value of measurement that depends on parameters \( \theta = (\theta_1, \ldots, \theta_p) \) we want to estimate: \( F(x_i; \theta) \)

Suppose that measurements are independent

To find best estimate of \( \theta \) we minimize the suitably weighted summ of squared differences between measured and predicted values → so called “least squares” or “chi-square”

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

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”)
    \[
    \chi^2(\theta) = \sum_{i=1}^{N} \left( \frac{(y_i - F(x_i; \theta))^2}{F(x_i; \theta)} \right)
    \]
    - But now \( \sigma \) depends on \( \theta \) which complicates the minimization
  - So called Neyman’s chi-square (or “modified chi-square”)
    \[
    \chi^2(\theta) = \sum_{i=1}^{N} \left( \frac{(y_i - F(x_i; \theta))^2}{y_i} \right)
    \]
- 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
  \[
  \frac{\partial \chi^2(\theta)}{\partial \theta_i} = 0, \quad i = 1, \ldots, p
  \]