

Statistics and Error and Data Analysis for Particle and Nuclear Physics

1 Introduction

Most students signing up for experimental physics courses envisage future work with complex devices, tools and electronics. On the other hand, students enlisting for theoretical physics studies typically think of days and nights behind the desk, working miracles with formulae. Very few realize that both experimental and theoretical achievements converge at one point: data analysis and hypothesis testing. In the fields of high energy and nuclear physics, most educational institutes do not host modern experimental facilities, thus data analysis is often the primary objective of doctoral theses for the experimentalists.

It has been observed that level of preparation of graduate students in the art of data analysis is unfortunately low. Often basic understanding of error treatment is missing, and knowledge of statistics is only preliminary, even though most students had some statistics courses during undergraduate studies. The reason is likely twofold: detachment of theory from actual practice, and underestimation of importance of these subjects in future research.

The goal of this course is to introduce the students to the specifics of data taking and further analysis in high energy and nuclear physics. Series of scenarios intend to demonstrate practical implications of principles of statistics, and to invite the students to do simple estimates themselves. Scenarios offer an introduction to key concepts of experiments and measurements, and an overview of basics of data and error analysis.

The course is worth 7.5 ECTS.

1.1 *Main Learning Objectives*

Main learning objectives of the course are:

- Learn specifics of a typical detector setup, how is it influenced by the physics requirements and what does it imply for data taking
- Learn basic procedures involved in measurement of physical observables
- Understand importance of proper data treatment in experiment
- Understand theory behind standard data treatment techniques
- Learn terminology used in relevant scientific literature and common practice
- Understand principles used in data processing and detector simulation software
- Learn how to work with relevant reference literature and online information sources

2 Scenarios

This compendium covers the following scenarios of the *Statistics and error analysis and data analysis for particle and nuclear physics (SED)* course:

- Introduction
- Principles of data correction
- Experimental errors
- Distributions
- Event reconstruction
- Understanding a sampling calorimeter

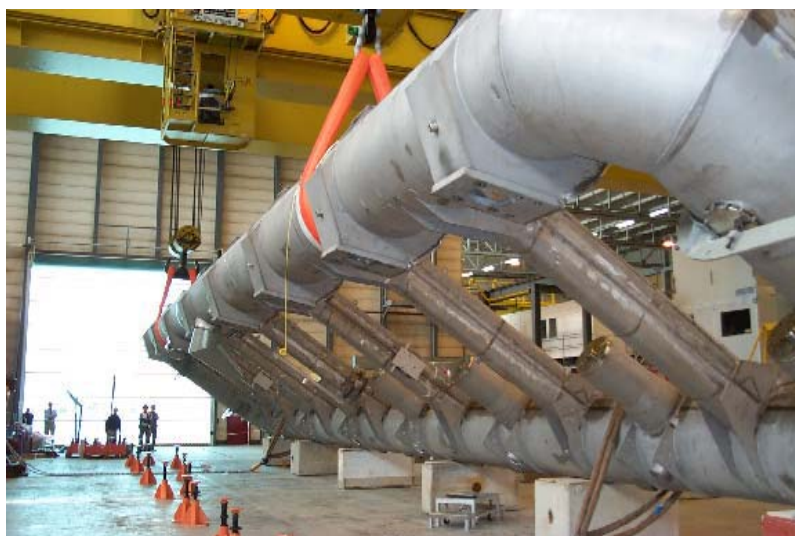
Each scenario is accompanied with the list of recommended literature and a text describing for supervisors the aims of the material.

Scenarios are organised in a sequential manner, starting with two introductory overviews, and continuing to more advanced scenarios concentrated on specific subjects. The first scenario bears specific marks of being one of the first PBL cycles offered to the group of students; hence it aims at engaging the students into active communication and into learning how to identify important questions for further studies. Following scenarios imply certain level of knowledge acquired by the students in due time. Every scenario introduces new terms that may be yet unknown to the students. Certain scenarios contain figures, for a better visual interpretation of the problem. Typically, scenarios do not foresee specific assignments except of the studies of the subject as such.

2.1 Introduction

2.1.1 Scenario: ATLAS Detector at LHC

On a rainy November afternoon, the security officer at CERN gates was surprised to see the expected delivery from Spain: a huge truck carrying 25-meter long, 5-meter wide construction of a round-cornered rectangle shape blocked the entire motorway, desperately trying to take turn and enter through the narrow gate. The students standing nearby were however not surprised: they started discussing enthusiastically that this must be the first of 8 *cryostats* that will hold the ATLAS *barrel toroid magnets*. This unique installation will provide a magnetic field of 4 Tesla, and will be one of the key components of a system that will facilitate efficient detection of decay products of awaited new particles, produced in collisions at the Large Hadron Collider.



2.1.2 Literature

1. ATLAS Detector and Physics Performance Technical Design Report, <http://atlas.web.cern.ch/Atlas/GROUPS/PHYSICS/TDR/access.html>
2. The ATLAS Experiment Web Site, <http://cern.ch/Atlas>
3. C. Grupen, Particle Detectors (Cambridge University Press, 1996).

2.1.3 Aims of the material

The primarily goal of this scenario is to introduce the students to the specifics of a modern particle physics detector, through reasoning behind its construction and purpose.

Another goal is to make the students more familiar with the PBL approach, such that this scenario can be used as one of the first cycles for a group of novice students. It aims at stimulating lively discussions and yet keeping the students focused on identifying further learning goals.

It may be useful to distribute in mid-cycle hand-outs with drawings of the detector, or to provide a poster with a detailed layout. Visual appreciation of such a complex structure is very important, especially for the students who had no prior encounter with analogous installations.

Ability to be able to present and defend publicly their work is crucial in the very competitive field of High Energy Physics. Communication skills are also very important for the rest of the PBL course. Moreover, students at this early stage are not yet fluent in specific English terminology, thus the teacher must create an opportunity for them to learn and to use the terms. Most efficient approach is to assign every student to a public talk in front of the group.

As an assignment, each student is asked to prepare a 15-20 minutes slide presentation covering a subsystem of the ATLAS detector. Students select presentation subjects themselves, filling their names in a hand-out table as shown below:

Sub-system	Name
Pixel detector	
Semiconductor Tracker (SCT)	
Transition Radiation Tracker (TRT)	
Electromagnetic Calorimeter (LArg)	
Hadron Calorimeter (Tile)	
Monitored Drift Tubes (MDT)	
Cathode Strip Chambers (CSC)	
Resistive Plate Chambers (RPC)	
Thin Gap Chambers (TGC)	

2.2 *Principles of data correction*

2.2.1 **Scenario: First encounter with real data**

John is a student with a particle physics group. His first task is to analyse angular distribution of certain particles produced in a collision experiment and fit it with a function $\sim \cos^2\theta$. John receives experimental data samples that provide high statistics, and software tools that can extract information about registered events and particles. After applying a standard algorithm that selects suitable events and particles, John plots the angular distribution, only to discover that it does not look like a smooth $\sim \cos^2\theta$ function at all. Instead, he sees a quite irregular distribution, with distinct gaps and other odd shapes. John's supervisor suggests that the simplest solution to the problem is to apply a “*correction factor*”, obtained with the help of “*Monte Carlo*”. Unfortunately, available Monte Carlo samples do not contain sufficient statistics, and John is advised to request more of those through the collaboration's production system. The system offers to produce a rather confusing variety of samples, labelled “*evgen*”, “*simul*”, “*digit*”, “*pileup*” and “*reco*”. After some investigation, John manages to submit the requests in the correct order, and in a while receives a sufficient amount of Monte Carlo samples, that allow him to evaluate a satisfactory correction factor and finally recover the expected experimental $\sim \cos^2\theta$ distribution.

2.2.2 Literature

1. G. Cowan, Statistical Data Analysis (Oxford Science Publications 1998)
2. ATLAS Detector and Physics Performance Technical Design Report,
<http://atlas.web.cern.ch/Atlas/GROUPS/PHYSICS/TDR/access.html>

2.2.3 Aims of the material

The goal of this scenario is to introduce the students to the fact that detectors have large inefficiencies, which must be taken into account while analysing the data. The scenario focuses on a rather simple approach that relies on Monte Carlo for evaluating the correction factor. The $\cos^2\theta$ observable is chosen not to illustrate a specific process, but to indicate a smooth featureless distribution, suitable for this correction procedure. For more advanced students, more demanding observables and more appropriate unfolding methods may be mentioned.

Another purpose of this material is to make the students familiar with specific terminology used to refer to various types of Monte Carlo-produced data. Such terminology is not necessarily the same across the experiments, but certain similarities are always present. Moreover, this offers a way to learn about different components stages in Monte Carlo and real data treatment, such as an original particle generation, detector simulation and reconstruction.

Reference literature in this area is rather scarce, as no common methodology exists.

2.3 *Experimental errors*

2.3.1 **Scenario: Good results with bad data**

Jane is preparing a hands-on exercise for physics students, and plans to make use of old data collected in electron-positron annihilation at LEP-I. Searching in the archives, she discovers that there are two data samples at her disposal: both come from leptonic Z-boson decays, one containing absolute momenta of e^- , and another – of e^+ . Unfortunately, due to corruption of the old disk, only a small part of the e^+ sample can be recovered, corresponding to just 4% of the e^- volume. Jane nevertheless decides to go ahead with the exercise, and makes a trivial estimate of Z-boson mass value. To her satisfaction, she finds that the *spread* of both distributions is about the same. With the small e^+ sample, she obtains the positron momentum value of 43.0 ± 10.0 GeV, and the larger e^- sample gives the corresponding momentum value of 45.0 GeV, with much smaller *error*. She ultimately estimates the Z mass at 89.8 ± 3.9 GeV. The students however come with a different result: 88.0 ± 10.2 GeV. Upon close inspection, Jane finds out that the students did not really commit any mistake: they simply did not make use of theoretical knowledge that would help reducing the error.

2.3.2 Literature

1. J. Orear, Notes on Statistics for Physicists (UCRL-8417, 1958)
2. D. J. Hudson, Lectures on Elementary Statistics and Probability (CERN 63-29, 1963)
3. W.T. Eadie et al, Statistical methods in experimental physics (North-Holland, 1971 and Elsevier, 1983)
4. S.L. Meyer, Data Analysis for Scientists and Engineers (Wiley, 1975)
5. S. Brandt, Statistical and computational methods in data analysis (North-Holland, 1976)
6. A. Frodesen et al, Probability and Statistics in Particle Physics (Oxford University Press, 1979)
7. L. Lyons, Statistics for Nuclear and Particle Physicists (Cambridge University Press 1986, reprinted 1989 and 1999)
8. R. Barlow, Statistics: A Guide to the Use of Statistical Methods in the Physical Sciences (Wiley, 1989)
9. B. Roe, Probability and Statistics in Experimental Physics (Springer, 1997)
10. J. Taylor, An Introduction to Error Analysis: The Study of Uncertainties in Physical Measurements (University Science Books, 1997)
11. G. Cowan, Statistical Data Analysis (Oxford Science Publications 1998)

2.3.3 Aims of the material

The goal of this material is to familiarize the students with the following aspects of error analysis:

- The meaning of sigma
- Combining errors
- Combining results from different experiments
- Using the error matrix

While not all of the mentioned items are covered by the specific scenario, the students will inevitably encounter them in the reference literature. It is a good idea to provide hints for students as to what to pay attention to when reading the books. It is also useful to mention that different sources may use different terminology, such as “uncertainty” versus “error” etc. In general, literature on this subject is abundant, and the teacher may chose to offer a different list.

2.4 *Distributions*

2.4.1 **Scenario: Working with histograms**

Paul studies journal publications about a specific track *observable* a , in order to find data on its measured distribution, dn/da . He finds an article that publishes this distribution as measured by an experiment during the first year of running. The data are presented in a table, showing for each a bin the value of dn/da normalized to total number of measured events and the corresponding error. As the observable is predicted to have a Gaussian distribution, Paul tries to *fit* it with a Gaussian shape, but uncertainties are too high. After some more search, he finds a newer article by the same experiment that publishes the same distribution, but measured during the second and third year of data taking. Luckily, both distributions are presented with the same binning, and despite they are normalized to a different total number of events, Paul is able to add up the two distributions after simply deducing the original number of entries in each bin, and the total number of events. The combined distribution has smaller uncertainties and produces a better overall fit.

2.4.2 Literature

1. J. Orear, Notes on Statistics for Physicists (UCRL-8417, 1958)
2. D. J. Hudson, Lectures on Elementary Statistics and Probability (CERN 63-29, 1963)
3. W.T. Eadie et al, Statistical methods in experimental physics (North-Holland, 1971 and Elsevier, 1983)
4. S.L. Meyer, Data Analysis for Scientists and Engineers (Wiley, 1975)
5. S. Brandt, Statistical and computational methods in data analysis (North-Holland, 1976)
6. A. Frodesen et al, Probability and Statistics in Particle Physics (Oxford University Press, 1979)
7. L. Lyons, Statistics for Nuclear and Particle Physicists (Cambridge University Press 1986, reprinted 1989 and 1999)
8. R. Barlow, Statistics: A Guide to the Use of Statistical Methods in the Physical Sciences (Wiley, 1989)
9. B. Roe, Probability and Statistics in Experimental Physics (Springer, 1997)
10. J. Taylor, An Introduction to Error Analysis: The Study of Uncertainties in Physical Measurements (University Science Books, 1997)
11. G. Cowan, Statistical Data Analysis (Oxford Science Publications 1998)

2.4.3 Aims of the material

The goal of this scenario is to familiarize students with the concept of histograms, which is the key tool in physics analysis. Understanding of histograms and manipulations with them requires good knowledge of the basics of statistics, and in particular of binomial, Poisson and Gaussian distributions. Though it is not explicitly suggested by the scenario, the students are expected on this stage to identify the necessity of re-visiting the basics of e.g. Poisson distribution. In case of difficulties in doing this, the teacher should provide the necessary hints.

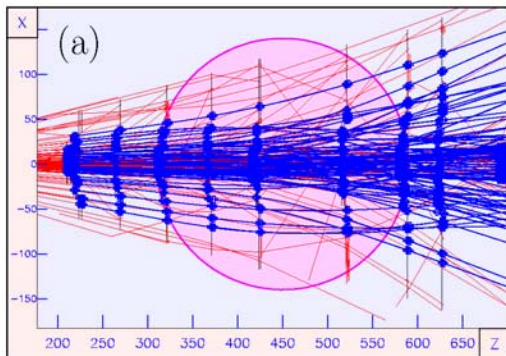
The students are not expected to study all the suggested reference literature, as practically every book offers a good description of the subject.

In the end, it could be considered an important achievement if every student will identify a favourite book or two on statistics, for further reference.

2.5 Event reconstruction

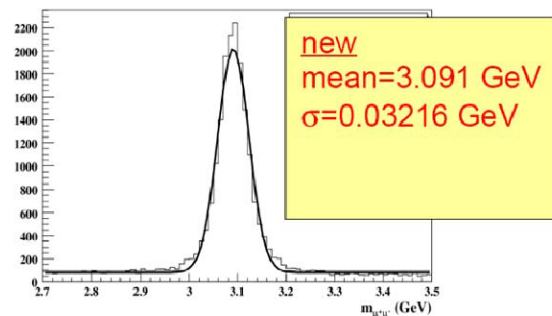
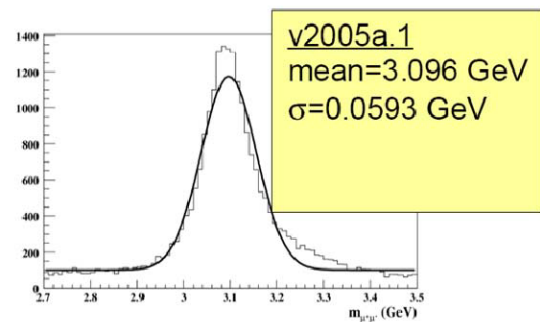
2.5.1 Scenario: Find a needle in a haystack

Peter starts as a student with a B-physics experiment, and his first task is to investigate why the measured J/ψ mass *resolution* is much worse than expected. After some studies, he starts to suspect problems with *event reconstruction*. He learns that reconstruction is a very complex process that involves many steps, starting with mapping electronic signals to actual geometrical location. Peter meticulously checks every reconstruction step, and discovers that the *calibration* database is outdated for one of the sub-detectors. Fixing this improves J/ψ mass resolution to $\sigma=0.06$ GeV, but Peter is still not satisfied. He reads an article describing an improved *track reconstruction* algorithm, and decides to give it a try. To his surprise, implementation of this algorithm brings J/ψ mass resolution to a much better value of $\sigma=0.03$ GeV, and the overall event reconstruction becomes much more satisfactory, even when all other reconstruction procedures and algorithms are kept unchanged. Peter describes all the reconstruction procedure and his contribution to its improvement in a thesis and defends it with an excellent grade.



Reconstruction of a golden B decay in the detector. Two tracks belong to muons from J/ψ decay

J/ψ mass
PDG = 3.096 GeV



2.5.2 Literature

1. C. Grupen, Particle Detectors (Cambridge University Press, 1996).
2. K. Kleinknecht, Detectors for Particle Radiation (Cambridge University Press, 1999).
3. D. Green, The Physics of Particle Detectors (Cambridge University Press, 2000).
4. R. Fruhwirth, M. Regler, R.K. Bock, H. Grote, D. Notz, Data Analysis Techniques for High Energy Physics, 2nd Edition (Cambridge University Press, 2000).
5. R. C. Fernow, Introduction to Experimental Particle Physics (Cambridge University Press, 1986).

2.5.3 Aims of the material

The aims of the material is to learn about stages involved in event reconstruction and how they affect final analysis, such as reconstructing invariant mass of a particle. Specific focus subjects include:

- Presentation of an event picture and list of measurement points
- Fitting the tracks
- Fitting the vertices
- Understanding the event
- Missing energy
- Statistical reconstruction of invariant mass

The presented scenario does not imply discussion on reconstruction of missing energy, but students are likely to come upon this issue during literature studies or even in the discussion, when properly guided. A teacher may chose to adjust the scenario such that this issue will become more pronounced.

The scenario is accompanied with illustrations in order to provide starting point for discussion. One may consider simpler figures with less occupancy.

2.6 *Understanding a sampling calorimeter*

2.6.1 Scenario: Optimizing detector design

A team of physicists is assigned to prepare a design proposal for an electromagnetic calorimeter for a modern high-energy physics experiment. The task is quite challenging, as calorimetry offers a wide range of possible combinations, unlike tracking. The team must achieve good energy and spatial resolution, and yet take into account physics requirements, detector size and cost limitations. The team decides to use *Geant4* detector simulation toolkit in order to test calorimeter models of various geometry and material composition. It turns out that the compromise between all the requirements can be achieved by constructing a *sampling calorimeter* with lead as an absorber and liquid argon as the active material. The concept is accepted, especially since it is further proposed to use a similar setup, only with copper instead of lead, to serve as a hadron calorimeter.

2.6.2 Literature

1. C. Grupen, Particle Detectors (Cambridge University Press, 1996).
2. K. Kleinknecht, Detectors for Particle Radiation (Cambridge University Press, 1999).
3. D. Green, The Physics of Particle Detectors (Cambridge University Press, 2000).
4. R. Fruhwirth, M. Regler, R.K. Bock, H. Grote, D. Notz, Data Analysis Techniques for High Energy Physics, 2nd Edition (Cambridge University Press, 2000).
5. R. C. Fernow, Introduction to Experimental Particle Physics (Cambridge University Press, 1986).
6. R. Wigmans, Calorimetry, Energy Measurement in Particle Physics (Oxford Scientific Publications, 2000).
7. Geant4: A toolkit for the simulation of the passage of particles through matter.
<http://cern.ch/geant4>

2.6.3 Aims of the material

The aim of this scenario is a closer study of one of the most important sub-systems of modern particle detectors, the sampling calorimeter. The other goal is to introduce the students to the Geant4 toolkit. More specifically, study subjects are:

- Geometrical description in Geant4
- Acceptance correction
- Response to electrons
- Factors affecting the resolution
- Response to charged pions

At this point, the group may already have knowledge of Geant4, and it is then conceivable to prepare an assignment involving simulation of a sampling calorimeter prototype. Otherwise, study of the reference literature and Geant4 usage manuals would be sufficient.