Modern Methods of Data Analysis

Introduction

WS 07/08 – Universität Heidelberg

Stephanie Hansmann-Menzemer
What do you learn in this Course?

The course goal is:

• Learn how to extract PHYSICS knowledge from measurements in (particle) physics

• Acquire competence in
  – understanding the statistical tools needed for data analysis
  – understanding the role of uncertainties and probabilities in relating experimental data and theory

• Are able to perform analysis of real data applying these techniques
What you NOT learn in this Course?

- How experimental detectors work.
- How the standard model is build.
- How to calculate Feynman diagrams.
- How to design and construct detector components.
- How to design large software projects.
- ...

To be a statistician is great! You never have to be “absolutely sure” of something ...
Being “reasonably certain” is enough!
(unknown)
The Particle Physics case ...

Why would one need to know “statistical methods for data analysis” in particle physics?

Why should I need to learn such methods?

Let's just consider the case of the Standard Model in particle physics ...
Components of the Standard Model

Materieteilchen

Leptonen

<table>
<thead>
<tr>
<th>Elektron</th>
<th>Myon</th>
<th>Tau</th>
</tr>
</thead>
<tbody>
<tr>
<td>e</td>
<td>μ</td>
<td>τ</td>
</tr>
<tr>
<td>\nu_e</td>
<td>\nu_μ</td>
<td>\nu_τ</td>
</tr>
</tbody>
</table>

Elektrische Ladung

-1  0  0

aufsteigende Masse

Quarks

<table>
<thead>
<tr>
<th>Up</th>
<th>Charm</th>
<th>Top</th>
</tr>
</thead>
<tbody>
<tr>
<td>u</td>
<td>c</td>
<td>t</td>
</tr>
<tr>
<td>d</td>
<td>s</td>
<td>b</td>
</tr>
</tbody>
</table>

Elektrische Ladung

2/3  -1/3

aufsteigende Masse

Kräfte & Austauschbosonen

EM Kraft

Photon γ

Starke Kraft

8 Gluonen

Schwache Kraft

W, Z Boson

+ Materieantiteilchen

+ Higgs
Experiments confirm standard model to incredible accuracy!

Everything is great - We found THE THEORY!

.. is this really all ???

Measurement of Z cross-section:
Data points, uncertainties smaller than symbol size, data confirm hypothesis of 3 light neutrino families
No! Some Questions remain ...

- What determines the mass of leptons and quarks?
- Why are the masses sooo different?
- Is there a unification of forces (GUT)?
- Can one integrate Gravitation into the SM?
- Why is there an asymmetry between matter and antimatter?
- Why does nature create several generations?
- ... many more ...

Need to improve in an interplay between theory and experiment
Methodology of Particle Physics

- Theory with Open Questions
  - Hypotheses
    - Predictions
  - Expand and Modify Hypotheses
  - Experiment, Measurement
  - Validate or discard Predictions

- Established Hypothesis
  - New Theory
How do we reach those goals?

- Introduction to statistical concepts in the lectures
- Hands-on work in the computer exercises
- End of semester: reproduce physics analysis on real data -> presentation
Outline of the lecture

- Basics Concepts & Definitions
- Characteristics of distribution
- MC generators
- Important distributions
- Error propagation
- Estimators
  - Maximum Likelihood
  - Least Square method
- Confidence Level and Limits
- Hypothesis Tests
- Add. Material
  - multi-variant systems, analysis bias, numerical methods
- Applied Examples
  - Alignment @ LHCb(J. Blouw), Kalman-Fitter/Bs mixing
Once upon a time, in a holiday resort the landlord L. ran a profitable B&B, and every morning bought 30 rolls for breakfast. By law the mass of a single roll was required to be 75 g. One fine day the owner of the bakery changed, and L. suspected that the new baker B. might be cheating. So he decided to check the mass of what he bought, using a kitchen scales with a resolution of 1g. After one month he had collected a fair amount of data:

73 79 72 62 67 60 60 67 78 68 66 75 76 73 75 64 70 69 73 59 70 73 64 72 64 69
69 71 69 71 77 69 72 71 67 72 63 66 68 76 71 76 68 71 63 65 65 66 73 73 73 67
70 65 71 69 78 67 65 69 71 71 72 73 72 69 66 66 70 60 72 62 53 65 74 65 68 69
67 75 64 76 72 76 78 67 67 67 69 79 71 67 71 68 71 65 66 65 78 76 71 70 67 65
67 64 73 67 74 79 74 71 73 67 66 76 68 74 76 65 77 67 71 67 71 77 63 66 70 62
68 74 67 67 67 77 65 68 79 72 71 77 68 70 73 67 81 70 74 71 79 62 67 63 68 76
73 81 76 73 68 72 76 61 69 73 71 80 68 70 62 76 58 68 68 64 68 78 69 65 70 70
64 75 73 72 60 86 68 68 64 60 68 71 70 75 70 67 69 67 73 65 66 71 70 70 73 66
72 71 71 64 76 75 72 72 71 72 71 75 68 73 70 64 76 72 75 79 70 64 70 67 70
75 70 83 69 61 70 66 69 71 72 70 76 73 62 71 60 73 74 70 68 68 70 78 71 69 71
73 73 75 65 71 67 60 70 77 71 74 64 74 73 60 77 73 70 69 66 70 78 69 75 66 71
75 75 74 69 74 70 75 77 75 66 72 68 72 61 75 65 69 68 65 73 82 67 75 67 80 71
79 72 71 68 73 70 67 75 74 69 63 63 72 70 73 63 70 70 59 78 76 66 72 79 65 71
76 72 69 69 73 70 77 73 83 66 68 67 69 73 76 65 71 70 71 65 78 71 67 70 72 75
67 79 72 64 62 79 68 70 61 65 68 71 73 60 60 68 71 74 75 69 73 70 68 ...
Data Reduction

- the raw list of number is not very useful
  - need some kind of data reduction
- assume that all measurements are equivalent
  - the sequence of individual data does not matter
  - all relevant information is contained in the number of counts per reading

```
count[50]= 0  count[60]= 20  count[70]= 85  count[80]=  9
count[51]= 0  count[61]= 11  count[71]= 81  count[81]=  7
count[54]= 0  count[64]= 31  count[74]= 54  count[84]=  0
count[55]= 0  count[65]= 48  count[75]= 43  count[85]=  0
count[56]= 2  count[66]= 42  count[76]= 33  count[86]=  1
count[57]= 1  count[67]= 70  count[77]= 23  count[87]=  0
count[58]= 3  count[68]= 68  count[78]= 21  count[88]=  0
```

- much improved presentation of the collected information
- the above numbers cover the entire data set
- most of the measurements are lower than 75g ...
Visualization

- an even better presentation of the available information
  - bar-chart of the tables given before
- example for the concept of a histogram
  - define bins for the possible values of a variable
  - plot the number of entries in each bin
    - immediate grasp of center and width of the distribution

The rolls produced by baker B. definitely show a lack of doe. So L. was right in his suspicion, that B tried to make some extra profit by cheating ....
and the Conclusion

As a consequence of his findings, L. complained to B. and even threatened to inform the police about his doings. B. apologized and claimed that the low mass of the rolls was an accident which will be corrected in the future. L., however, continues to monitor the quality delivered by the baker. One month later, B. inquired again about the quality of his products, asking whether now everything was all right. L., for his part, acknowledged that the weight of the rolls now matched his expectations, but he also voiced the opinion that B. was still cheating ...

the histogram shows: B. simply selected the heaviest rolls for L. !
Literature

- Statistische und numerische Methoden der Datenanalyse (V. Blobel/E. Lohrmann)
- Statistical Data Analysis (Glen Cowan)
- Statistics: A guide to the use of statistical methods in Physical Sciences (R. J. Barlow)

Recommendation for a rainy Sunday afternoon:
Many good lectures on statistical methods around

  Lecture SS07 Christoph Grab & Christian Regenfus Uni Zuerich

- Michael Feindt, Guenther Quast, IEKP Karlsruhe, many different lectures and many good exercise examples!

- Guenter Duckeck, Madjid Boutemeur  Lecture 7.10. - 11.10.2002
  http://www-alt.physik.uni-muenchen.de/kurs/Computing/stat/stat.html

- Ian C. Brock, WS 2000/01, Universitaet Bonn
  http://www-zeus.physik.uni-bonn.de/~brock/teaching/stat_ws0001/index

- Michael Schmelling, Heidelberger Graduiertentag 2006

Got lot's of inspiration from their courses - Thanks!
Questionnaire:

• Which semester are you?
  – 5, ≥7, Diplomand, Doktorand
  – Knowledge on particle physics?
• Did you attended statistic lecture before?
  – school, graduate courses, ...
• Programming skills?
  – C++, ROOT, other languages
• Do you plan to attend computer course?
  – Do you have a Rechenzentrum-account?
• Please write down your email.
Modern Methods of Data Analysis

Lecture I (15.10.07)

Contents:

• Basic concepts
• Definitions ...
Why do we bother?

If your experiment needs statistics, you ought to have done a better experiment!
(B. Russell)

The result of this experiment was so inconclusive, so we had to use statistics!
(overheard at international physics conference L. Lyons)
Why do we need statistics ...

Remember:

“Prediction is very difficult, especially if it's about the future!”
(N.Bohr)

But, anyway:

“It is far better to foresee even without certainty than not to foresee at all.”
(Poincare)
Predictable

For easy classical physical processes the result is precisely determined
$\rightarrow$ determinism “Determinisums”

Examples are:

pendular, orbit of planets, billard ...
Contingency

Random events are as a matter of principle not predictable (even with precise knowledge of the starting conditions)

Examples are:

► Lotto (depend on many quantities, deterministic chaos)
► radioactive decay (quantum mechanics)
► electronic noise
► measurement uncertainties
Quantum mechanics:
Each time something different happens ....

events of the CDF experiment
Measurements:

Experiments measure frequency distributions:

![Graph showing mass distribution with statistical uncertainties and number of candidates.](image-url)
Probability vs. Statistics

• Probability: from theory to data
  – Start with a well-defined problem and calculate all possible outcomes of a specific experiment

• Statistics: from data to theory
  – Try to solve the inverse problem: starting from a data set, want to deduce the rules/laws
    → this is analyzing experimental data
  – deals with parameter estimation
    • determine parameters and uncertainties in unbiased and efficient way
    • hypothesis testing (agreement, confidence ...)

Stephanie Hansmann-Menzemer
Interpretation of Probability

- mathematical probability (axiomatic)
- frequentist probability:
  - objective probability definition
  - interpretation/definition as relative frequency
- subjective probability (Bayesian)

Beware of HOW you apply and interpret different results!

Beware of different “names” used in literature!

Lucky enough, in (most) the useful cases they use the same combinatorial rules of probabilities.
Kolmogorov Axioms (1931)

Elementary events $e_i$ are considered, which exclude each other:

- $e_i$ = elementary event
- $\Omega$ = set of all elementary events
- $P(e_i) =$ probability of event $e_i$

1. $P(e_i) \geq 0$ for all $i$
2. $P(e_i \lor e_j) = P(e_i) + P(e_j)$ ($\lor =$ or)
3. $\sum_{\Omega} P(e_i) = 1$

Formal approach, however from pragmatic point of view rather useless ...
Objective Probability

- “Objective prior” probability:
  - define as a hard, objective quantity, a property of the experiment (symmetry of a coin, dice ...)
- Outcome of experiment can happen in \( n \) different ways with equal probability (throwing dice, flipping a coin ...) and \( k \) of the outcomes have a certain property \( A \):
  \[ P(A) = \frac{k}{n} \]

Disadvantage: Definition does not hold for continuous variables ...
Examples:

Daughters & Sons
A woman has two kids. She is asked: “Do you have a daughter” and her answer is “yes”. What is the probability that her second child is a girl?

Quiz show in American TV:
One of three doors hides a car (all three equally likely), and the other two hide goats. You choose Door 1. The quiz master, who knows where the car is, then opens one of the other two doors to reveal a goat, and asks whether you wish to switch your choice. Say he opens Door 3; should your stick with your original choice, Door 1, or switch to Door 2?
Frequentist Probability

- Empirical definition via frequency of occurrence (R. von Mises)
  - perform experiment $N$ times in identical trials, assume event $E$ occurs $k$ times, then: $P(E) = \lim_{N \to \infty} \frac{k}{N}$
  - intuitive definition for particle physics (many repetitions)

- very useful, but has some problems
  - limes in strict mathematical sense does not exist, i.e. can not be proven that it converges
    ... how large is $N$? When does it converge?
  - What means repeatable under identical conditions? Is similar OK as well? ... this IS difficult, if not impossible ...
  - Not applicable for single events:

  e.g. “It will probably rain tomorrow!” there's only ONE tomorrow, we wait and see. BUT, can do this only ONCE. So what does it mean?
Subjective (Bayesian) Probability

Reverend Thomas Bayes (1702-1761)

Probability is the degree of believe, that an experiment has a specific result.

Subjective probability compatible with Kolmogorov axioms

Essay Towards Solving a Problem in the Doctrine of Chances (1763), published postum in Philosophical Transactions of the Royal Society of London
Examples for Bayesian Probability

Frequentists interpretation often not applicable. Than Bayesian interpretation only possible one.

Probability is the degree of belief that a hypothesis is true:
  - The particle in this event is a positron.
  - Nature is supersymmetric.
  - It will rain tomorrow.
  - Germany will win the soccer championship 2010.

Often criticized as “subjective” and “non scientific”. However it is based on simple probability computation (axioms). It is not contradicting the Frequentists approach, but include it.
Bayes' Theorem (1)

Conditional (“bedingte”) probability

\[
P(B|A) = \frac{P(B \cap A)}{P(A)} \quad \text{Due to } P(A \cap B) = P(B \cap A)
\]

\[
P(A|B) = \frac{P(A \cap B)}{P(B)} = \frac{P(B|A) P(A)}{P(B)}
\]
Bayes' Theorem (2)

Important through the interpretation $A=\text{theory}, B=\text{data}$

$$P(\text{theory}|\text{data}) = \frac{P(\text{data}|\text{theory})P(\text{theory})}{P(\text{data})}$$
Examples: Rare Disease (1)

Probability of disease A:  \( P(A) = 0.001 \)
\[ P(\text{not } A) = 0.999 \]

Test for disease:

\[
\begin{align*}
P(+|A) &= 0.98 \\
P(-|A) &= 0.02 \\
P(+|\text{not } A) &= 0.03 \\
P(-|\text{not } A) &= 0.97 
\end{align*}
\]

Do you need to be worried if you get “+” as test result?
What is the posteriori probability?
Example: Rare Disease (2)

\[
P(A|+) = \frac{P(+|A)P(A)}{P(+|A)P(A) + P(+|noA)P(noA)}
\]

\[
= \frac{0.98 \times 0.001}{0.98 \times 0.001 + 0.03 \times 0.999}
\]

\[
= 0.032
\]

The posterior probability is only 3.2%, due to the tiny prior probability and the non negligible misidentification rate.
Bayesians vs. Frequentists

Someone studying large sample of potential carriers of disease:

Prior probability: overall fraction of people who carry disease
Posterior probability: fraction of people who are carriers out of those with positive test result

A specific individual, however, may be interested in the subjective probability:

Prior probability: Degree of belief that one has the disease before the test
Posterior probability: Degree of belief that one has the disease after the test

Frequentist point of view: Probability that individual has disease is 0 or 1, we just don't know. Bayesians say it is 0.1%.
Bayesians vs. Frequentists

There are strong rivaling schools of these approaches.

A frequentist is a person whose lifetime ambition is to be wrong 5% of the time.

A Bayesian is one who, vaguely expecting a horse, and catching a glimpse of a donkey, strongly believes he has seen a mule.

Looks for the moment like philosophical question, difference in interpretation will come clearer once we discuss confidence levels & statistical and systematical uncertainties. Again: Math is the same!
Combination of Probabilities

Can be deduced from Kolmogorov axioms:

\[
\begin{align*}
P(A \text{ or } B) &= P(A) + P(B) - P(A \text{ and } B) \\
P(A \text{ and } B) &= 0 \\
P(A \text{ or } \bar{A}) &= P(A) + P(\bar{A}) = 1 \\
P(A \text{ and } B) &= P(A) \cdot P(B|A) = P(B) \cdot P(A|B) \\
P(A \text{ and } B) &= P(A) \cdot P(B)
\end{align*}
\]
Example RICH detector to use Bayes' theorem:

- Proton-Antiproton beam:
  90% pions, 10% kaons
- kaon ID 95% efficient, pion mis-ID 6%

Question: If the RICH indicates a kaon, what is the probability that it is a real kaon/a real pion?

\[
P(A|B) = \frac{P(B|A) \times P(A)}{P(B|A) \times P(A) + P(B|\text{not } A) \times [1 - P(A)]}
\]
Non informative prior

If you nothing about a quantity (even not order of magnitude) uniform distribution is wrong assumption!

\[ f(x) = \text{const} \quad \text{WRONG!} \]

Rather \[ f(\ln(x)) = \text{const} \quad \text{RICHTIG!} \]

corresponds to \[ f(x) \sim 1/x \]

Bradfordsche's law of numbers: The digit 1 is way more often the significant digit of a number than 9 (independent of system). BTW: The tax office is aware of it, be careful by cheeting tax declarations.
Two more things ...

- Tomorrow Computer Exercises:
  - Getting started with ROOT

Slides of the lecture available:

http://www.physi.uni-heidelberg.de/~menzemer/statistik_vorl.html