WS 2022/2023

Freitags 14:00 – 17:00
CIP Pool, INF 226 1.305

Einführung in die Datenanalyse mit dem C++ Toolkit ROOT

Jörg Marks, Physikalisches Institut, INF 226
marks@physi.uni-heidelberg.de

Programm Überblick

- Mikrowiederholung Linux und C++ Konzepte
- Einführung in die Datenanalyse mit dem Analysewerkzeug ROOT

Organisatorisches

- 2 Leistungspunkte:
  - Anwesenheitspflicht mit Lösung der Übungsaufgaben
  - Vortrag
- Kurs web page

https://www.physi.uni-heidelberg.de/~marks/root_einfuehrung/

Informationen zur Veranstaltung (1)

Ziele

- C++ Mikrowiederholung und Ergänzungen
  - Um die C++ Schnittstelle von ROOT effektiv nutzen zu können.
  - ROOT Quellcode ansehen zu könn en.
  - https://www.physi.uni-heidelberg.de/~marks/c++_einfuehrung/
  - Pythonintegration
- Einführung in die Datenanalyse unter Verwendung des ROOT Toolkits
  - Input / Output von Messungen und Resultaten
  - Graphische Darstellung von Messungen
  - Statistische Methoden der Datenauswertung
  - Datenanpassung zur Bestimmung von Modellparametern mit Minuit und rooFit
  - Multivariate Datenanalyse
- Beispielorientiert Konzepte so erläutern, dass Sie mit den Erklärungen selbständig (kleine) Datenanalyseaufgaben lösen können.
  - Tutorial Stil
  - Grundlagen für das Erstellen problemorientierter Lösungen schaffen

Voraussetzungen

- C++ Vorkenntnisse notwendig, hohe Informationsdichte und Tempo
- User ID zur Benutzung der CIP Pools der Fakultät für Physik

Informationen zur Veranstaltung (2)

Struktur des Kurses

- Wechsel zwischen Vorlesung und Übungen
- Wechsel zwischen selbstständigem Üben und Übungen in Kleingruppen
- Erläutern und Diskutieren der Lösungsvorschläge
- Kurszeiten: Freitags 14:00 – 17:00 (4 stündige Veranstaltung)
- Kurs Web Page:
  http://www.physi.uni-heidelberg.de/~marks/root_einfuehrung/

  - Vorlesungstransparente
  - Beispiel Code
  - Übungsaufgaben
  - Lösungsvorschläge

Voraussetzungen für einen Leistungsnachweis (2 LP)

- Anwesenheitsliste / mehr als 1 x Abwesenheit nur mit Attest
- Aktive Mitarbeit und kleine Übungen als Hausarbeit
- Die Klausur wird durch einen Seminarvortrag ersetzt
- Keine Benotung

Einleitung und Motivation

Daten

- Norm des internationalen Technologiestandards (ISO/IEC 2382-1, 1993)
  data: „ a reinterpretable representation of information in a formalized manner, suitable for communication, interpretation, or processing “
- Informatik
  - Maschinenlesbare und -bearbeitbare, digitale Repräsentation von Information.
  - In Zeichen bzw. Zeichenketten kodiert, deren Aufbau Regeln (Syntax) folgt.
  - Um aus Daten wieder die Informationen zu abstrahieren, müssen sie in einem Bedeutungskontext interpretiert werden.
- Speicherung der Daten auf Festplatten, Magnetbändern, Flashspeicher, ...
  - Zahl der Internet Nutzer 2022: $5 \cdot 10^9$ [91% (85%) der Bevölkerung USA (EU)]
  - Erwartete jährliche Datenmenge 2025: $175 \cdot 10^{12}$ GBytes (Faktor 6 zu 2018)
  - Globaler IP trafic 2022: $396 \cdot 10^9$ GBytes / Monat (Faktor 2.5 zu 2018)
- Vorlesung: durch Messung / Beobachtung gewonnene Information in Form von Zahlen und Text mit folgenden Eigenschaften
  - die Menge der Daten ist typischer Weise groß, keine Durchsicht „von Hand“
  - die gemessenen Werte müssen in physikalische Information verwandelt werden
  - die Information liegt nicht in reiner Form vor, sondern ist in anderen Daten versteckt

  → Lerne Techniken zur Verarbeitung von Daten / Informationsextraktion

Einleitung und Motivation

Datenanalyse

- Individuelle Lösung durch Erstellen eines selbstgeschriebenen Computerprogramms zur Interpretation (Auswertung) der Daten

  - z. B. Mittelwerte, Zeitabhängigkeit, graphische Darstellung, Anpassung und Extraktion von Modellparametern, …

    → ok , aber nicht sehr effizient

- Verwende Toolkit, das möglichst viel von der Programmierarbeit vornimmt

  - vorgefertigte Programmbausteine
  - Beispiele: mathematica, matlab, origin, …..
  - Nachteil: proprietäre Software mit eigener Syntax (Programmierinterface), (obwohl häufig auch eine Anbindung an höhere Programmiersprachen existiert)

    → Es lassen sich nur sehr schwer eigene Anpassungen vornehmen.

    → Kein Zugang zu den verwendeten Algorithmen.

- Vorlesung: verwende Toolkit, das Public Domain Software ist und ein interpretiertes Sprachinterface zu C++ und Python und zum C++ Compiler hat.

  ROOT Toolkit, das zur Analyse der LHC Daten entwickelt wurde / wird.

Messungen und Messfehler

Messung

$$x_m = x_w + \Delta x$$

x_m : gemessener (angezeigter) Wert der Messgrösse

x_w : wahrer Wert der Messgrösse (nicht bekannt)

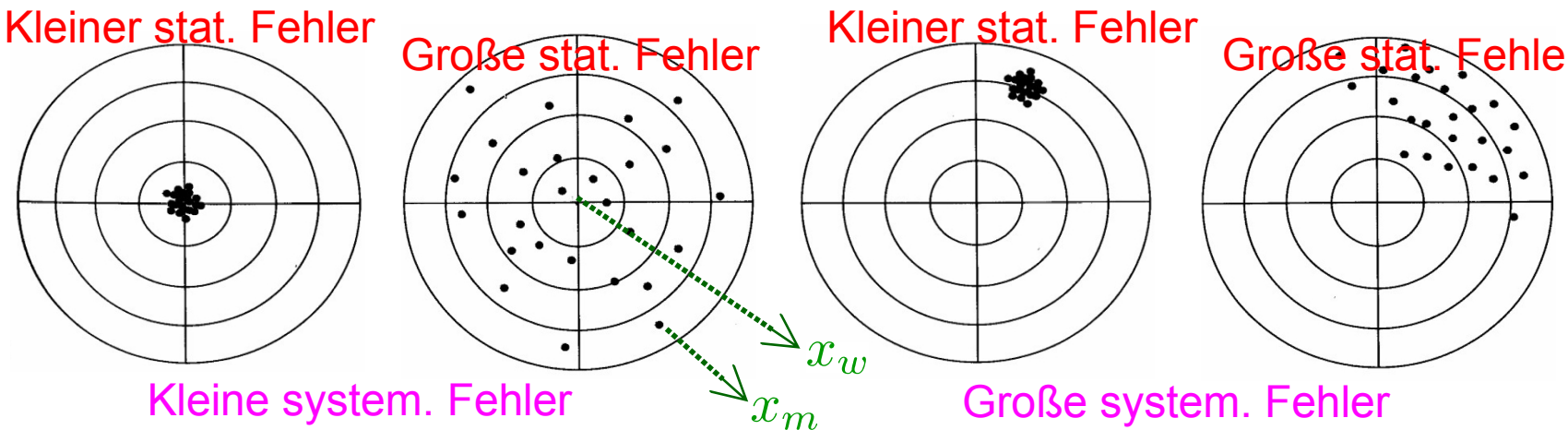
Δx : Messabweichung (Messfehler)

Messfehler

Eine Messung erfolgt immer nur mit endlicher Genauigkeit, 2 Beiträge:

- Systematische Fehler: Konstante, einseitg gerichtete Abweichung vom wahren Wert unter gleichen Messbedingungen.
- Zufällige oder statistische Fehler: Zufällige, nicht einseitig gerichtete Abweichungen vom wahren Wert (Mittelwert M und Messunsicherheit s).

Beispiel: Messergebnisse einer Sternposition



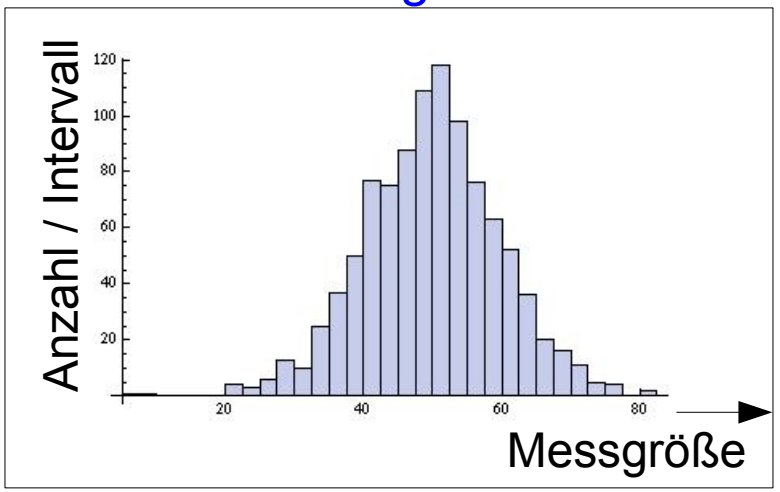
Fehlerrechnung

Quantitative Bestimmung

- Systematische Fehler:
  Schwierig! Genaue Analyse des Messaufbaues
- Zufällige oder statistische Fehler:
  Mehrfache Messung der selben Größe
- Schätzung des Messwertes bei mehrfacher Messung:

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$$ Arithmetrisches Mittel

Histogramm



- Fehler einer Einzelmessung

  Eigenschaft des Arithmetischen Mittels $\sum_{i=1}^{n}(x_i - \bar{x})^2 = min$

$$\sigma_E = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}}$$ Standardabweichnung

- Mittlerer Fehler des Mittelwertes
  Mittelwert von n Messungen ist um $\frac{1}{\sqrt{n}}$ genauer als die Einzelmessung

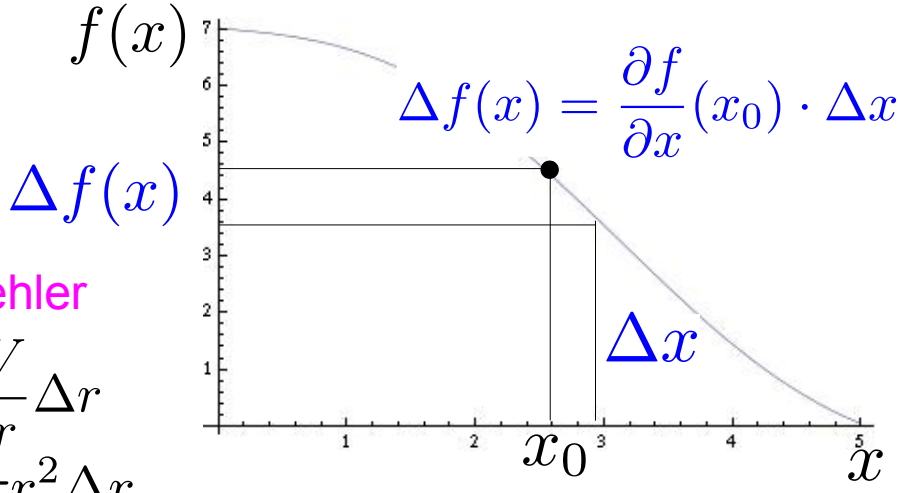
$$\sigma_M = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n(n-1)}}$$

Fehlerfortpflanzung

Fehler zusammengesetzter Größen

Wie wirkt sich ein gemessener Fehler auf eine zusammengesetzte Größe aus?

$$\Delta f(x) = \frac{\partial f}{\partial x}(x_0)\Delta x$$



Beispiel: Fehler von V bei gegebenem Fehler von r

$$V = \frac{4}{3}\pi r^3 \qquad \Delta V = \frac{\partial V}{\partial r}\Delta r$$

$$\Delta V = 4\pi r^2 \Delta r$$

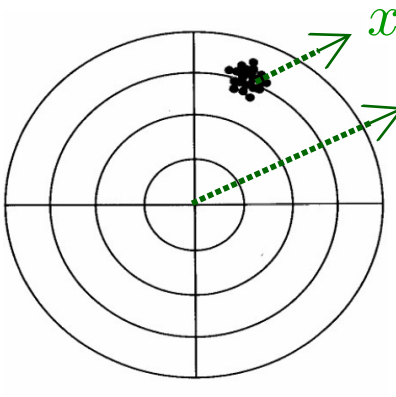
Gauß'sches Fehlerfortpflanzungsgesetz

Wie wirken sich mehrere gemessene Fehler auf eine zusammengesetzte Größe aus?

$$\kappa = f(x, y) \qquad \Delta\kappa = \sqrt{(\frac{\partial f}{\partial x}\Delta x)^2 + (\frac{\partial f}{\partial y}\Delta y)^2}$$

Analyse der Messungen

Messung Sternposition



Anforderungen an ein Computer-Programm zur Analyse von gemessenen Daten:

- Messdaten einlesen, x_m
- Kalibration / Berechnungen $x = \frac{1}{N}\sum_i^N x_i - x_w$
- Datenanpassung / Bestimmung von Modellparametern mit Fehlern

1 D Gauss Funktion:

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

- Graphische Darstellung
- Simulationsrechnungen / theoretische Beschreibung und Vergleich mit den Messungen
- Ausgabe der Ergebnisse

Messwerte mit 2 D Gauss: