Statistische Methoden der Datenanalyse

Ulrich Husemann

Humboldt-Universität zu Berlin
Wintersemester 2010/2011

Vorstellung

- Vorlesung: Ulrich Husemann
  - Nachwuchsgruppenleiter bei DESY und HU Berlin seit 2008
  - Arbeitsgebiet: experimentelle Teilchenphysik, ATLAS-Experiment am Large Hadron Collider (CERN)
  - E-Mail: ulrich.husemann@desy.de (immer)
  - Büro DESY: Platanenallee 6, Zeuthen, Raum 3L/27, Tel.: 033762-7-7392 (meistens)
  - Büro Adlershof: NEW 15, 2'412 (nach Absprache)
- Übung: Valentina Ferrara
  - E-Mail: valentina.ferrara@desy.de
  - Büro Zeuthen: Raum 2L/15, Tel.: 033762-7-7312

Einordnung

- Teil des Moduls P23.1.2b: Vorlesung aus der Reihe „Aktuelle Probleme der experimentellen Teilchenphysik"
- Zielgruppe:
  - Monomaster Physik, Spezialisierungsfach Elementarteilchenphysik
  - Geeignet für alle Masterstudierende sowie für Bachelorstudierende ab dem 3. Semester
- Nützliches Vorwissen: Grundlagen der Analysis und der linearen Algebra
- Praktischer Nutzen im Studium: Praktikum, Bachelor-/Masterarbeit (nicht nur experimentelle Teilchenphysik)

Termine & Arbeitsleistungen

- Termine: 2 VL + 1 Ü
  - Vorlesung: Donnerstags, 11–13 Uhr c.t., NEW 15 3'101
  - Übung: Montags, 15–17 Uhr c.t., NEW 15 1'427 (jede zweite Woche)
- 5 Studienpunkte durch „regelmäßige aktive Teilnahme an den Übungen":
  - Präsens in Übungen: mindestens 4 der 7 Veranstaltungen
  - Bearbeitung von Übungsaufgaben: mindestens 50% der erreichbaren Punkte
  - Vorrechnen von Übungsaufgaben oder Vorführen von Computerprogrammen: mindestens einmal

Übungen

- Übungsgruppenleiterin: Valentina Ferrara
- Übungstermine:
  - 14-täglich, erster Termin: Montag, 25.10.10, 15 Uhr c.t.
  - Weitere Termine bis Weihnachten: 01.11., 15.11., 29.11., 13.12.
  - Ort: Computer-Pool NEW 15 4'127
- Alle 14 Tage: Übungszettel
  - „Rechenaufgaben" zur Datenanalyse
  - Entwurf von Computerprogrammen
  - Ausgabe und Abgabe der Rechenaufgaben: Donnerstags in/nach der Vorlesung

Ablauf der Übungen

- „Praktische Übungen"
  - Implementierung von Problemen der Datenanalyse am Computer mit ROOT-Programmpaket
  - Natürlich auch: Besprechung der „Rechenaufgaben", Fragen zur Vorlesung
- Was ist ROOT (http://root.cern.ch)?
  - C++-Programmbibliothek und interaktives Programm (C++-Interpreter, auch Interfaces zu Python, Ruby)
  - DER Standard für Datenanalyse in der Teilchenphysik
  - Frei erhältlich für viele Computerplattformen
  - Erste Übung am 25.10.: kurze Einführung in ROOT

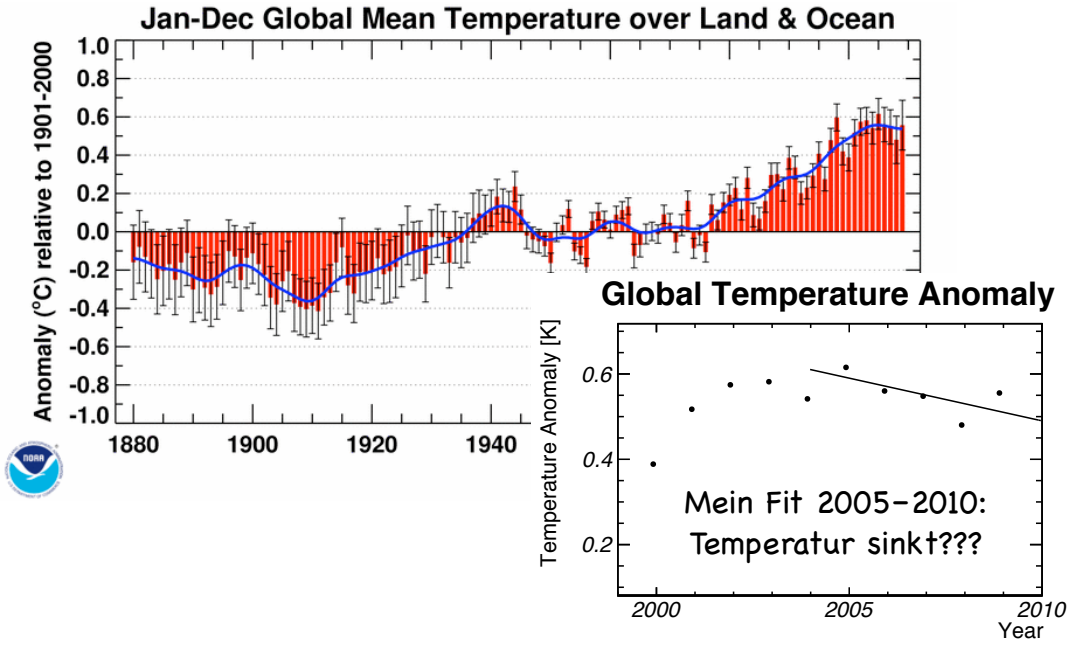
Literatur

- G. Cowan, Statistical Data Analysis, Oxford (1997)
- R. J. Barlow, Statistics: A Guide to the Use of Statistical Methods in the Physical Sciences, Wiley (1989)
- S. Brandt, Datenanalyse, Spektrum (1999)
- V. Blobel, E. Lohrmann, Statistisiche und numerische Methoden der Datenanalyse, Teubner (1998)
- G. Bohm, G. Zech, Introduction to Statistics and Data Analysis for Physicists, DESY E-Buch (2010)
- F. James, Statistical Methods in Experimental Physics, World Scientific (2006)

Motivation

- Sammlung und Analyse von Daten
  - Forschung in Naturwissenschaften, Medizin
  - Finanzwelt: Börsendaten, Wechselkurse, ...
  - Data-Mining in der Wirtschaft: Google, Payback-Karten, ...
- Test von Hypothesen, Klassifizierung von Daten, Bewertung von Risiken
  - Wurde am LHC ein neues Teilchen gefunden?
  - Ist diese E-Mail Spam?
  - Gibt es eine globale Erwärmung?
- Als Naturwissenschaftler sollten Sie solche Vorgänge verstehen und bewerten können!

Beispiel: Klimaforschung



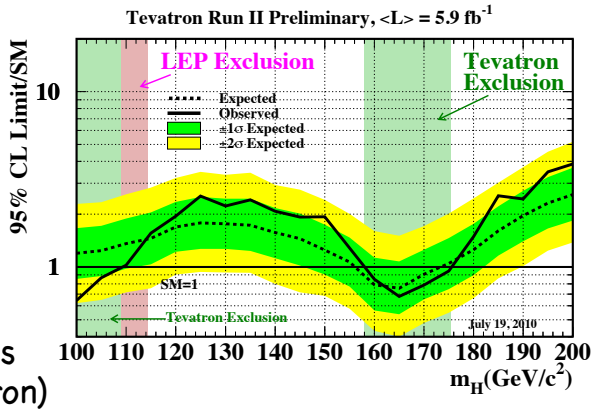
[http://www.ncdc.noaa.gov/cmb-faq/anomalies.html]

Konkretere Fragestellungen

- Auswertung von Messreihen:
  - Schätzung von Parametern (z.B. Mittelwert)
  - Verteilung der Messwerte um den Mittelwert
  - Anpassung von Funktionen (z.B. „Ausgleichsgerade")
  - Kombination von Messungen
- Unsicherheit von Messgrößen:
  - Angabe von Messwert ohne Unsicherheit ist sinnlos!
  - Statistische und systematische Unsicherheiten, z.B. Masse des Top-Quarks: m_t = 173.3 ± 0.6 (stat.) ± 0.9 (syst.) GeV/c^2

Konkretere Fragestellungen

- Weiterführende Auswertung
  - Wahrscheinlichkeit für Auftreten von Ereignissen
  - Signifikanz eines Messsignals (z.B. im Vergleich zum Rauschen)
  - Entscheidung über Modellhypothesen aufgrund der Messung (z.B. Ausschluss von Higgs-Massen am Tevatron)

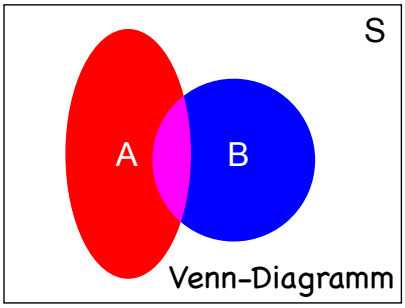


  - Entfaltung einer „wahren" Verteilung unter Berücksichtigung von Auflösungseffekten (z.B. Schärfen von digitalen Bildern)
  - Statistische Klassifizierung von Daten (z.B. mit neuronalen Netzen)

Inhalte und Ziele der VL

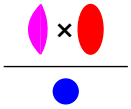
- Grundlagen der statistischen Datenanalyse für Physiker, praktische Beispiele
- Vorläufige Inhaltsangabe:
  - Grundlagen der Statistik
  - Wahrscheinlichkeitsverteilungen
  - Stichproben
  - Monte-Carlo-Methoden
  - Parameterschätzung
  - Prüfung von Hypothesen
  - Statistische Klassifizierung

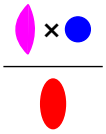
Satz von Bayes in Bildern

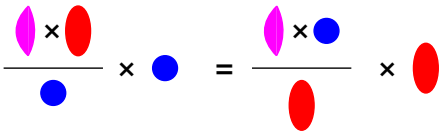


$P(A) =$ 

$P(B) =$ 

$P(A|B) =$ 

$P(B|A) =$ 



$$P(A|B) \times P(B) = P(B|A) \times P(A)$$

Thomas Bayes
(1702–1761)