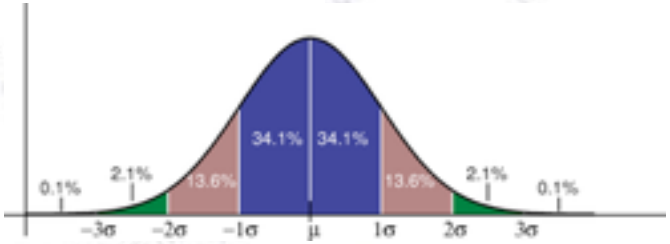
Applied Statistics

Course information






Troels C. Petersen (NBI)



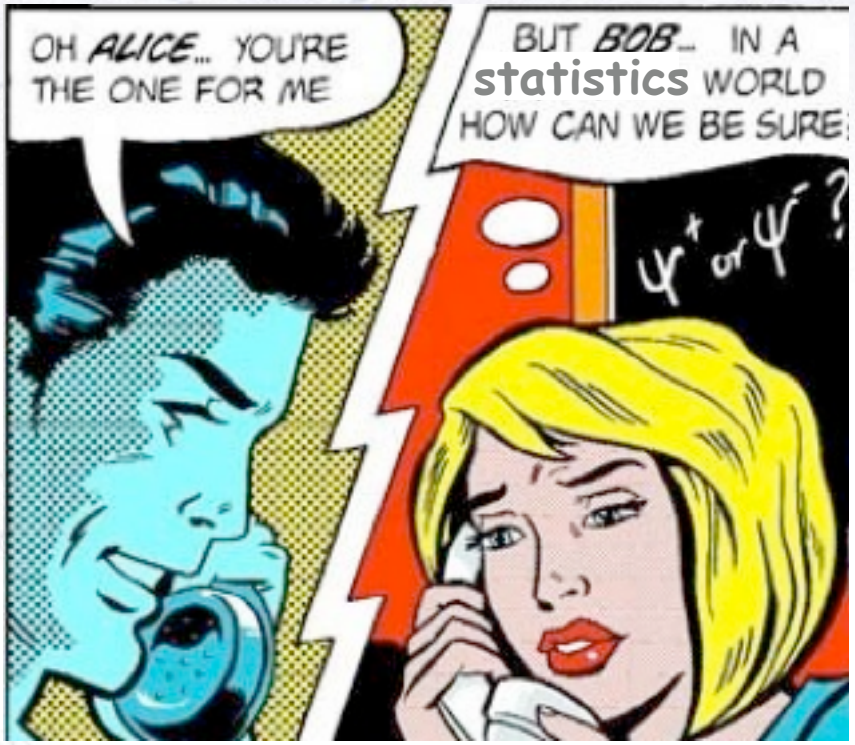
"Statistics is merely a quantisation of common sense!"

Applied Statistics 2015

…all the technical stuff!

Technicals:

- Rooms and hours.
- Course structure and dates.
- Computers and software.
- Data sets.
- Literature.
- Curriculum.
- Problem set.
- Projects.
- Exam.
- Expectations.
- Goals.



http://www.nbi.dk/~petersen/Teaching/AppliedStatistics2015.html


My office
(top floor)
Frimurerlogen
Entrance to Auditorium M
Entrance to Auditorium A
Blegdamsvej

Hours and Rooms

Hours:

Following block B, but using the morning hours 8:15 - 9:00 Monday and Friday for "self-studying".

Rooms:

Lectures: Auditorium M (except first day)
Exercises: Auditorium M
We also have the room next to Aud. M.

Monday:
9:15 - 10:00 Lectures (Aud. M)
10:15 - 12:00 Exercises (Aud. M)

Tuesday:
13:15 - 14:00 Lectures (Aud. M)
14:15 - 17:00 Exercises (Aud. M)

Friday:
9:15 - 10:00 Lectures (Aud. M)
10:15 - 12:00 Exercises (Aud. M)

First week: Python/ROOT intro Tuesday 13:15-14:00.

Hours and

Hours:

Following block

morning

10

First Tuesday 13:15-14:00.

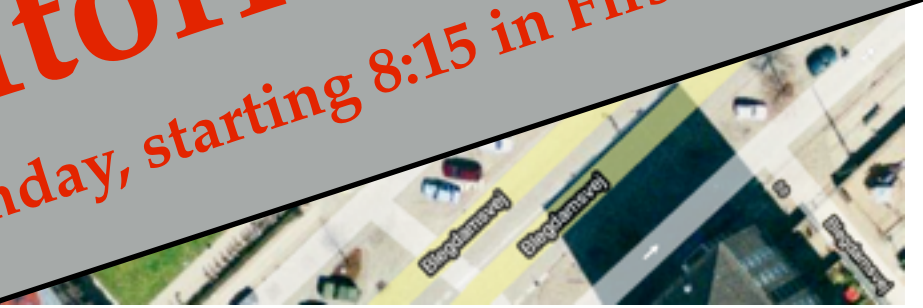
Only exceptions:

First day of course

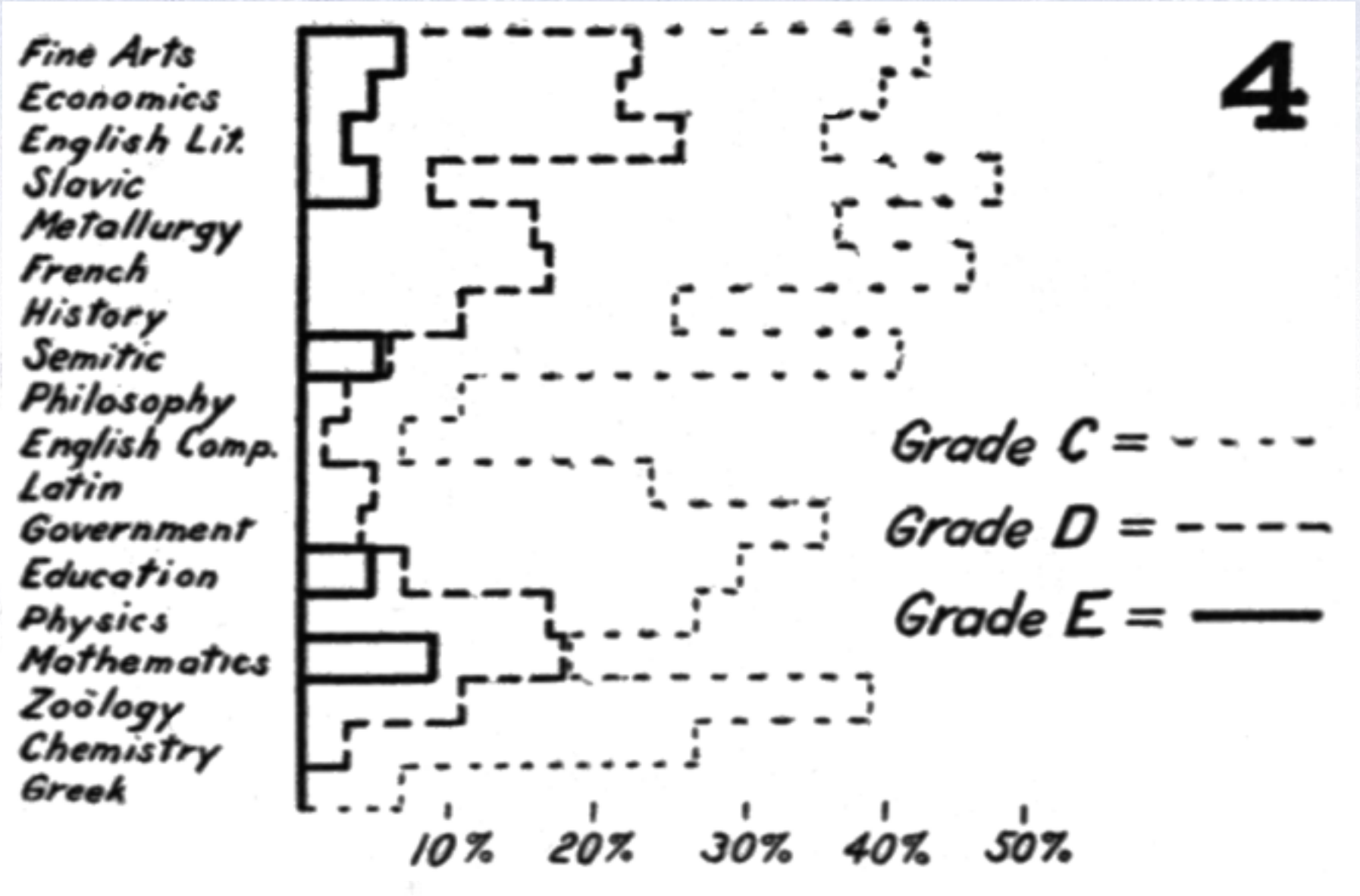
16th of November 8:15

in Auditorium A

...and the second Monday, starting 8:15 in First Lab



Computers and software



Computers and software

The times are ***way past*** pencil and/or calculator stage!
Fast computers is the ***only*** answer to do (any serious) data analysis.

Operating system: **Linux/MAC OS/Windows**
Editor: **Emacs** (or your own favorit!)
Programming: **Python** - version 2.7.X (not 3.0).
Higher level analysis: **ROOT** - version 5.34.X (not 6.0).

Installation:
http://www.nbi.dk/~petersen/Teaching/Stat2015/installation_python_root.html

Before course start, we will give an introduction to this ("Week 0"):
Thursday 12th 10:15-12:00 in Aud. M: Help with **installation**.
13:15-15:00 in Aud. M: Introduction to **programming**.

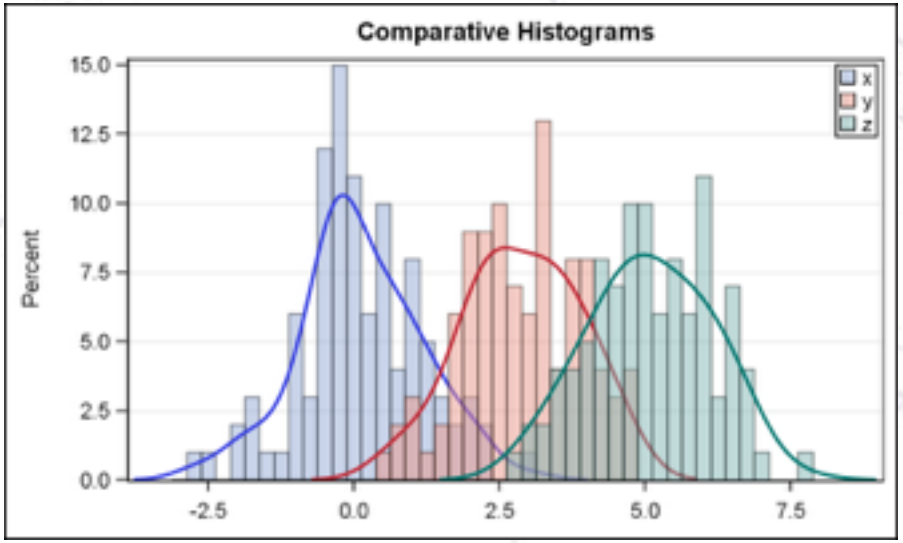
Also, during first week, first hour of Tuesday and possibly Friday after class, we will further introduce and train programming.

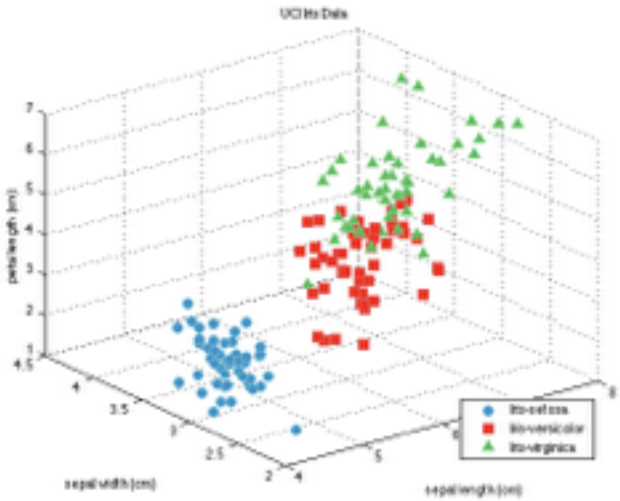
Data sets

In general, any data set can be used for this course! If you happen to have an interesting and illustrative one, bring it to me/class!

I've tried my best to search for a large variety of data sets, but this is not always easy. Publicly available data sets are often old/small/biased/etc.

As a result, some data sets are from my own field (particle physics). This is both due to my access to data here, but also because particle physics is one of the only fields providing ***billions of measurements***.





Literature

I chose to use Roger J. Barlow's "Statistics", as it is the best overall book.

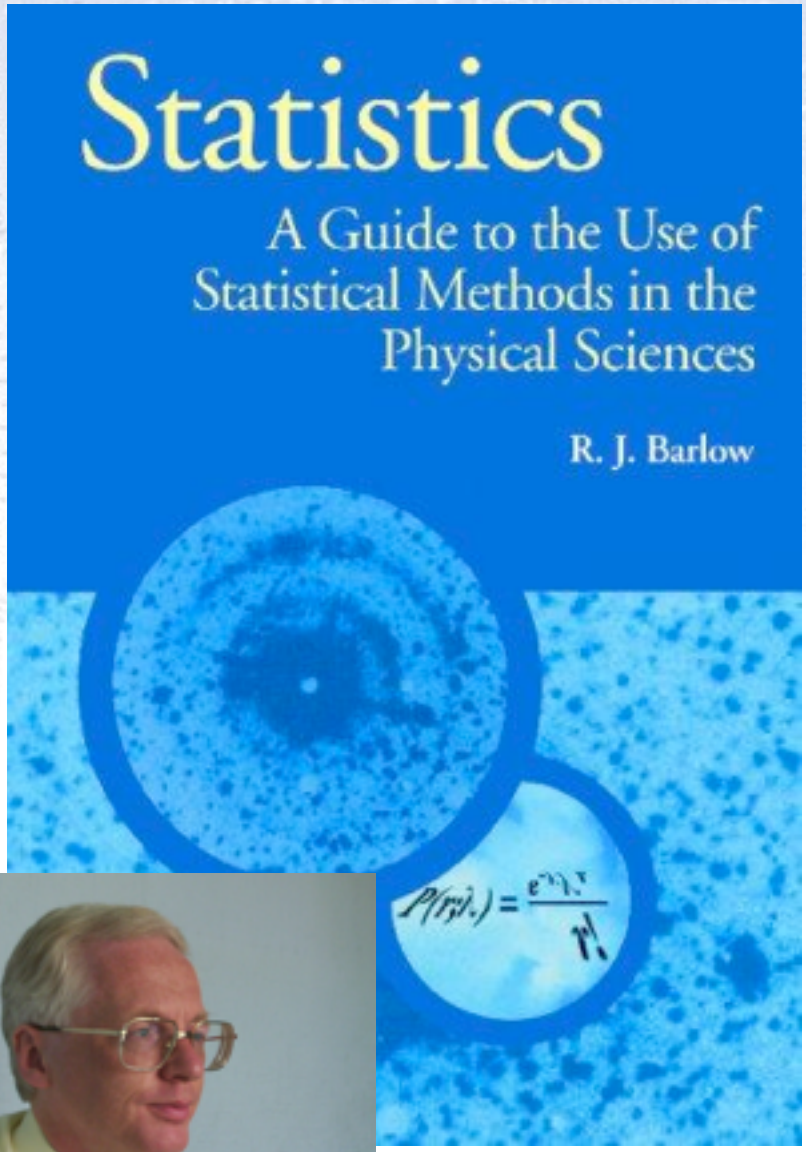
It is a very good and accessible introduction to statistics, and it gives many examples.

If anything, it is lacking a bit on how to generate random numbers according to a specific PDF and on categorizing events.

I will occasionally also refer to:

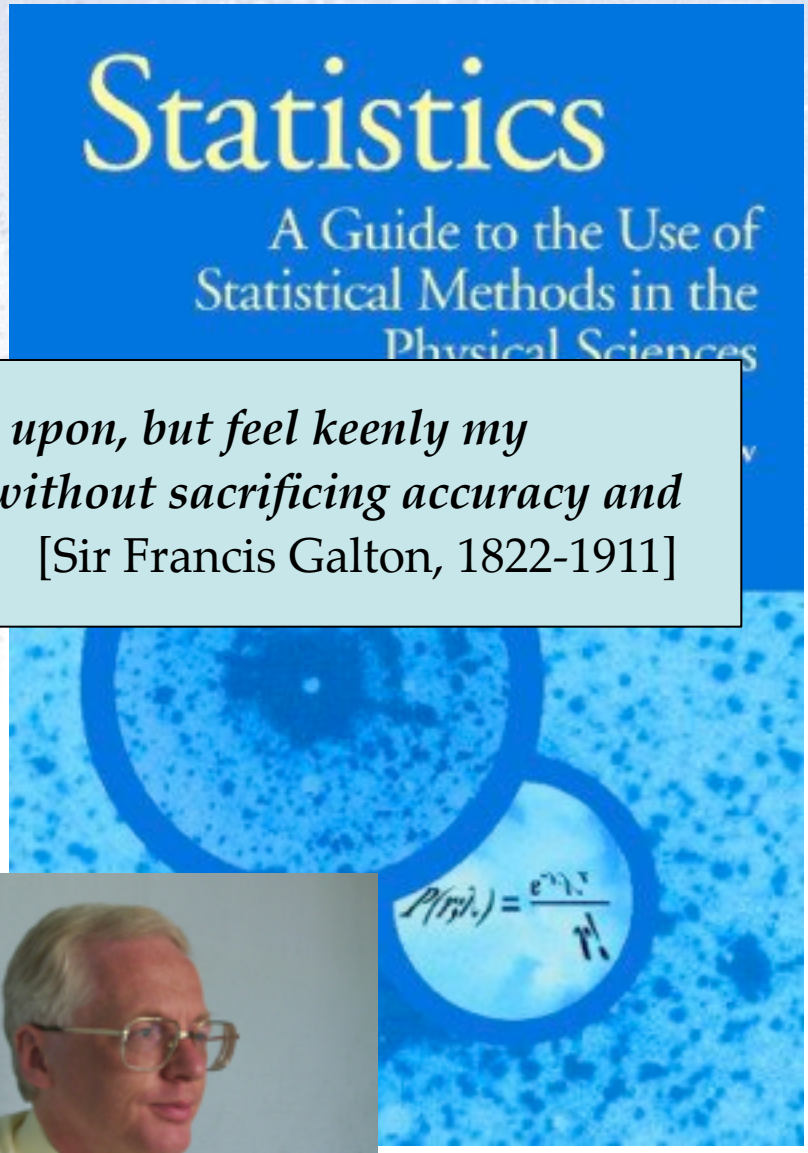
- A. Bevington:
  Data Reduction and Error Analysis
- Glen Cowan:
  Introduction to Statistics

...and notes from Particle Data Group!



Literature

I chose to use Roger J. Barlow's "Statistics", as it is the best overall book.



It is a very good and accessible introduction to statistics, and it gives many examples.

If
ge
sp

> ***"I have a great subject [statistics] to write upon, but feel keenly my literary incapacity to make it intelligible without sacrificing accuracy and thoroughness"*** [Sir Francis Galton, 1822-1911]

I will occasionally also refer to:

- A. Bevington:
  Data Reduction and Error Analysis
- Glen Cowan:
  Introduction to Statistics

...and notes from Particle Data Group!

Curriculum

The course will cover the following chapters in R. Barlow:

- Chapter 1 (All)
- Chapter 2 (All)
  Exercises: All, except 2.5 and 2.9.
- Chapter 3 (Except 3.2.2, 3.3.2, 3.4.2, 3.5.2)
  Exercises: All, except 3.7.
- Chapter 4 (All)
  Exercises: All, except 4.10.
- Chapter 5 (Except 5.1.3, 5.3.2, 5.3.3 (formal part), 5.3.4, 5.5)
  Exercises: 5.2
- Chapter 6 (Except 6.4.1, 6.7)
  Exercises: All
- Chapter 7 (Except 7.3.1)
  Exercises: All, except 7.1, 7.3, and 7.7.
- Chapter 8 (Except 8.4.4, 8.4.5, 8.5.1, and 8.5.2)
  Exercises: All, except 8.6.
- Chapter 10 (All)

Core of Curriculum

The course will **focus mostly on** the following chapters in R. Barlow:

- Chapter 2: 2.1, 2.2, 2.3, 2.4.1, 2.4.2, 2.6
- Chapter 3: 3.1, 3.2, 3.2.1, 3.3, 3.3.1, 3.4.1, 3.4.7, 3.5.1
- Chapter 4: 4.1, 4.2, 4.3, 4.3.1, 4.3.2, 4.3.3
- Chapter 5: 5.1, 5.1.1, 5.1.2, 5.2, 5.6
- Chapter 6; 6.1, 6.2, 6.2.1, 6.2.2, 6.2.3, 6.2.4, 6.3, 6.4
- Chapter 8: 8.1, 8.2, 8.3, 8.4, 8.4.1, 8.4.2, 8.4.3

This is less than 80 pages, but... they do not only require reading!
They request understanding!!!

The plan is to go through this curriculum in 4-5 weeks, spending the rest of the time on applying it.

It is through application that statistics is really understood.

Check list

In order for me to consider you inscribed in this course, you should fulfill the following check list *(within the first week!)*:

- **Have read the course information** (these slides, on course webpage).
  Otherwise, you don't know what is going to happen.
- **Have your picture ("mug shot") taken.**
  Otherwise, I don't know who you are.
- **Have filled in the questionaire** (on course webpage).
  Otherwise, I don't know what you know and don't.
- **Have measured the length of the lecture table in Auditorium A.**
  Otherwise, you haven't contributed to a common course dataset.
- **Have Python and ROOT running on your laptop**.
  Otherwise, you can't follow the exercises or solve problems.
- **Have registered for exam!**
  Otherwise, the administration will kill us!

In order not to continuously be doing the above, we will be doing all of the steps today and Friday after class, only!

Problem set

During the course (week 3-4), I will give a larger problem set to be solved and handed in.

This will cover most of the curriculum covered at this point, and it ***will count 15% in your final grade!!!***

It will require the use of computers and modifications of some of the code you have been running.

You are welcome to work in groups, but each student must hand in their own solution.

The final exam will somewhat resemble this problem set!

Projects

During the course (week 2-3 and week 6-8) you will be working on a larger data analysis project for about 1-2 week(s).

Each of these is your chance to play with real data and gain experience of what planning an experiment and detailed data analysis requires!
These ***will count 25% in your final grade!!!***

They will require the use of computers and modifications of some of the code you have been running.

You are encouraged to work in groups, and only one report (2-4 pages) is required from each group.

Real life problems will resemble these projects very much!

Projects

Project 1:

Attempt at precision measurement of the Earth's gravitation locally at NBI, using only "simple" means (OK - a little bit of cheating there).

You will be doing two separate experiments (both seen before):

- Ball rolling down an incline.
- Simple pendulum.

Project 2:

Whatever you would like to do, as long as it involves (advanced) data analysis! However, be warned that getting data is hard! So start today, or ask me for data known to work (reasonably well), such as:

- Gravity measurements and influence of Moon and Sun.
- Alpha decay measurements.
- Lifetime of the K_{short} particle with ATLAS/LHC 2015 data.
- Data from a company that needs analysing.
- Fun data from others…

Projects

Project 1:

Attempt at precision measurement of the Earth's gravitation locally at NBI using only "simple" means (OK - a little bit of cheating there).

You will be doing two separate experiments (both seen before):

- Ball rolling down an incline.
- Simple pendulum.

Project 2:

Whatever you would like to do, as long as it involves (advanced) data analysis! However, be warned that getting data is hard! So start today, or ask me for data known to work (reasonably well), such as:

- Gravity measurements and influence of Moon and Sun.
- Alpha decay measurements.
- Lifetime of the K_{short} particle with ATLAS/LHC 2015 data.
- Data from a company that needs analysing.
- Fun data from others…

Project 2 starts today!!!

Exam

Exam will be a 28 hour take-home exam with a problem set, which resembles the one previously given.

It will cover most of the curriculum, and it *will count 60% in your final grade!!!*

It will require the use of computers and modifications of some of the code you have been running.

You must work on your own!

I will provide this exam on:
Thursday the 21st of January 8:15am.

It will then naturally have to be handed in:
Friday the 22nd of January before 12:00!

Challenges

During the course, there will be a few challenges:

- Best table measurement analysis.
- Most precise measure of g (to better than 1/10000?).
- A problem on the problem set.
- Project 2
- ???

They are meant as advanced exercises to those, who are not already challenged significantly by the course! They do not give credit, but will of course earn you advanced experience and impress me (who gives grades and might be writing you a letter of recommendation).

Don’t stress over this - you can of course still earn the grade 12 without ever touching upon them.

Expectations

I want (read: insist) this course to be useful to all of you!

Therefore, please give me feedback (during the course, thanks!) if you have anything to add/suggest/criticise/alter.

However, it is mostly through your active participation that you have this privilege (i.e. that I'll listen more).

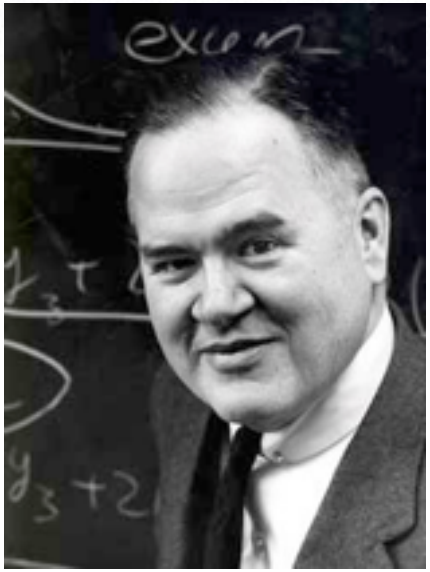
This also means, that I will require much from you - as much as I can without spoiling the social life of your youth!

In return, I'll try to make statistics as interesting as possible (and not deprive you of your early mornings).

Statistical practices

The famous statistician John Tukey (1915-2000) was quoted for wanting to teach:

- The **usefulness and limitation of statistics**.
- The importance of having methods of statistical analysis that are robust to violations of the assumptions underlying their use.
- The need to amass experience of the behavior of specific methods of analysis in order to provide guidance on their use.
- The importance of allowing the possibility of data's influencing the choice of method by which they are analysed.
- The need for statisticians to reject the role of "guardian of proven truth", and to resist attempts to provide once-for-all solutions and tidy over-unifications of the subject.
- **The iterative nature of data analysis.**
- Implications of the increasing power, availability and cheapness of **computing facilities**.
- The training of statisticians.



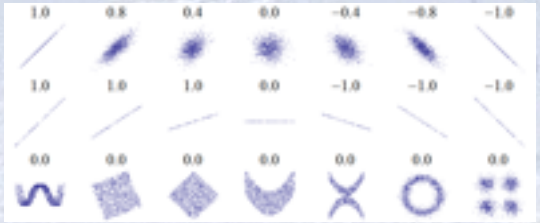
"Far better an approximate answer to the right question, which is often vague, than an exact answer to the wrong question, which can always be made precise." J. W. Tukey

References

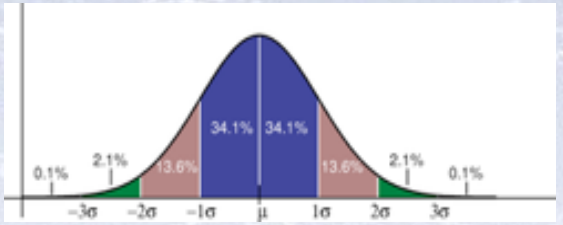
Philip R. Bevington: Data reduction and error analysis.
Classic introduction with very good examples - a standard reference in all of experimental physics [and so essentially all of physics!].

Roger J. Barlow: Statistics *(course book!)*
(A guide to the use of statistics methods in the physical sciences)
Very good introduction, which goes a little further than Bevington. Very much to the point.

Glen Cowan: Statistical Data Analysis
A bit brief, but once you got the hang of statistics, this book contains much of what you will ever need.



Top 10



Most important things in applied statistics

1. Errors decrease with the **square root of N**
2. **ChiSquare** is simple, powerful, robust and provides a **fit quality** measure
3. **Binomial** distribution → **Poisson** distribution → **Gaussian** distribution
4. **Error propagation** is **craftsmanship** - **fitting** is an **art**
5. Error on a (Poisson) number, N: $\sqrt{N}$ on a fraction, f=n/N: $\sqrt{f(1-f)/N}$.
6. **Correlations** are important and needs consideration
7. Hypothesis testing of H_0 (null) and H_1 (alt.) is done with a test statistic t
8. The **likelihood** (ratio) is generally the optimal estimator (test)
9. Low statistics is terrible – needs special attention
10. Prior probabilities needs attention, i.e. Bayes' Theorem