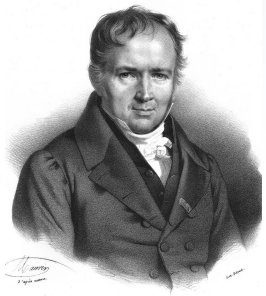
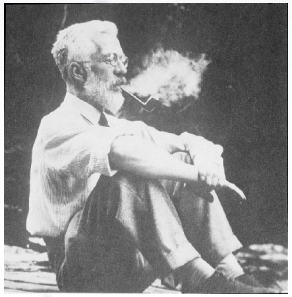
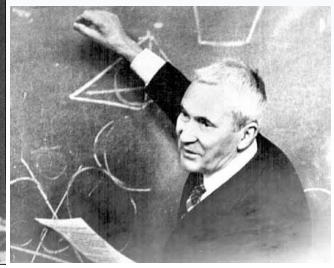
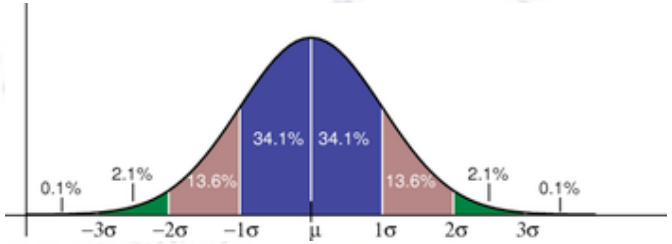
Applied Statistics

Course information 2019-20



Troels C. Petersen (NBI)



"Statistics is merely a quantisation of common sense!"

Applied Statistics 2019

…all the technical stuff!

Technicals:

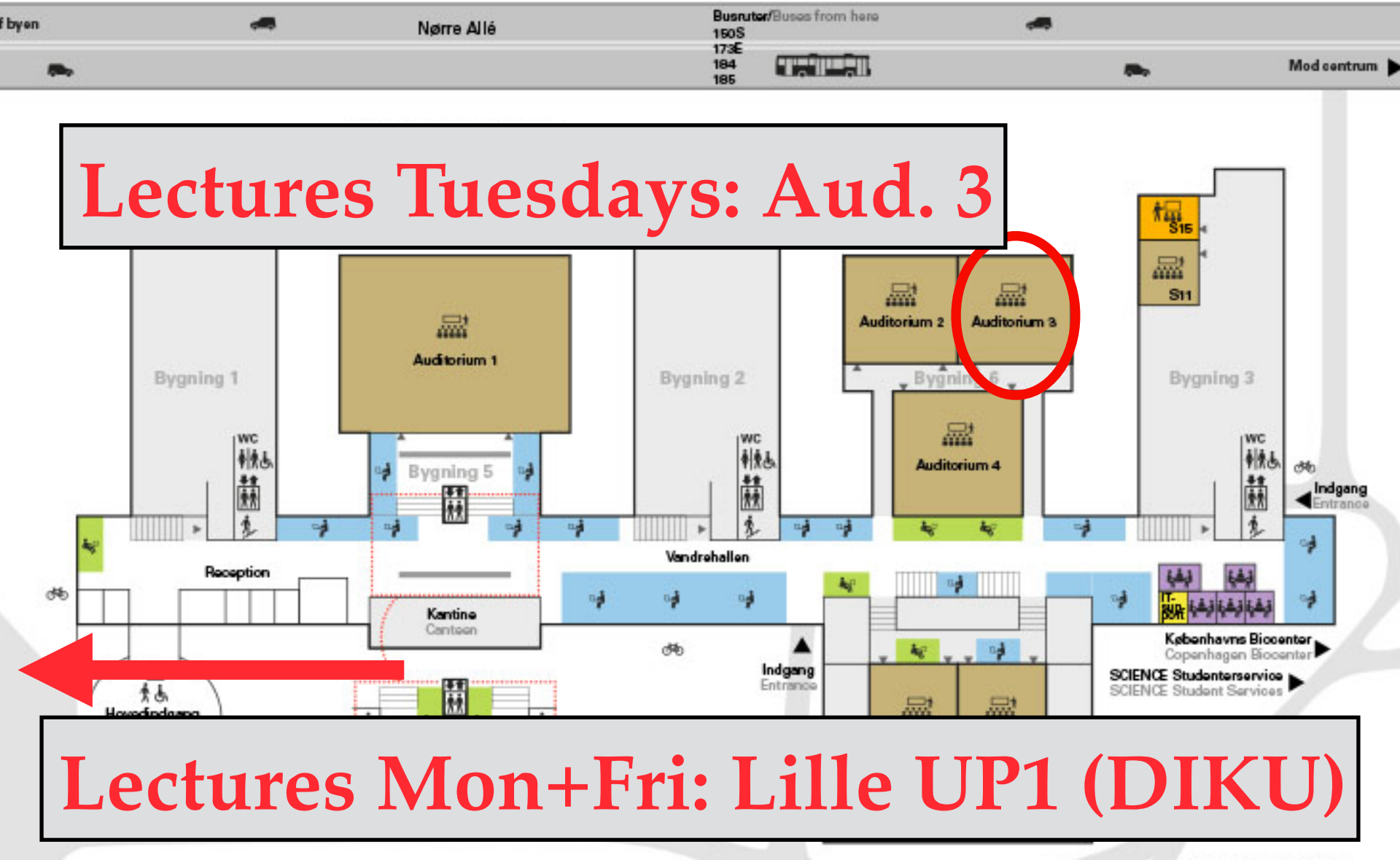
- Rooms and hours.
- Course structure and dates.
- Computers and software.
- Data sets.
- Literature.
- Curriculum.
- Problem set.
- Projects.
- Exam.
- Expectations.
- Goals.

The course webpage (central source of course information, bookmark or fail!):

http://www.nbi.dk/~petersen/Teaching/AppliedStatistics2019.html

Click on link in PDF, as copying text might not correctly get the "~" character right (especially on Windows!)

Lectures at HCØ & DIKU



Lectures at HCØ

Only exceptions:

First day of course

18th of November 8:15

in Auditorium A

...and the 2nd Friday/3rd Monday, starting 8:15 in First Lab

Le on+Fri: Lille UP1 (DIKU)

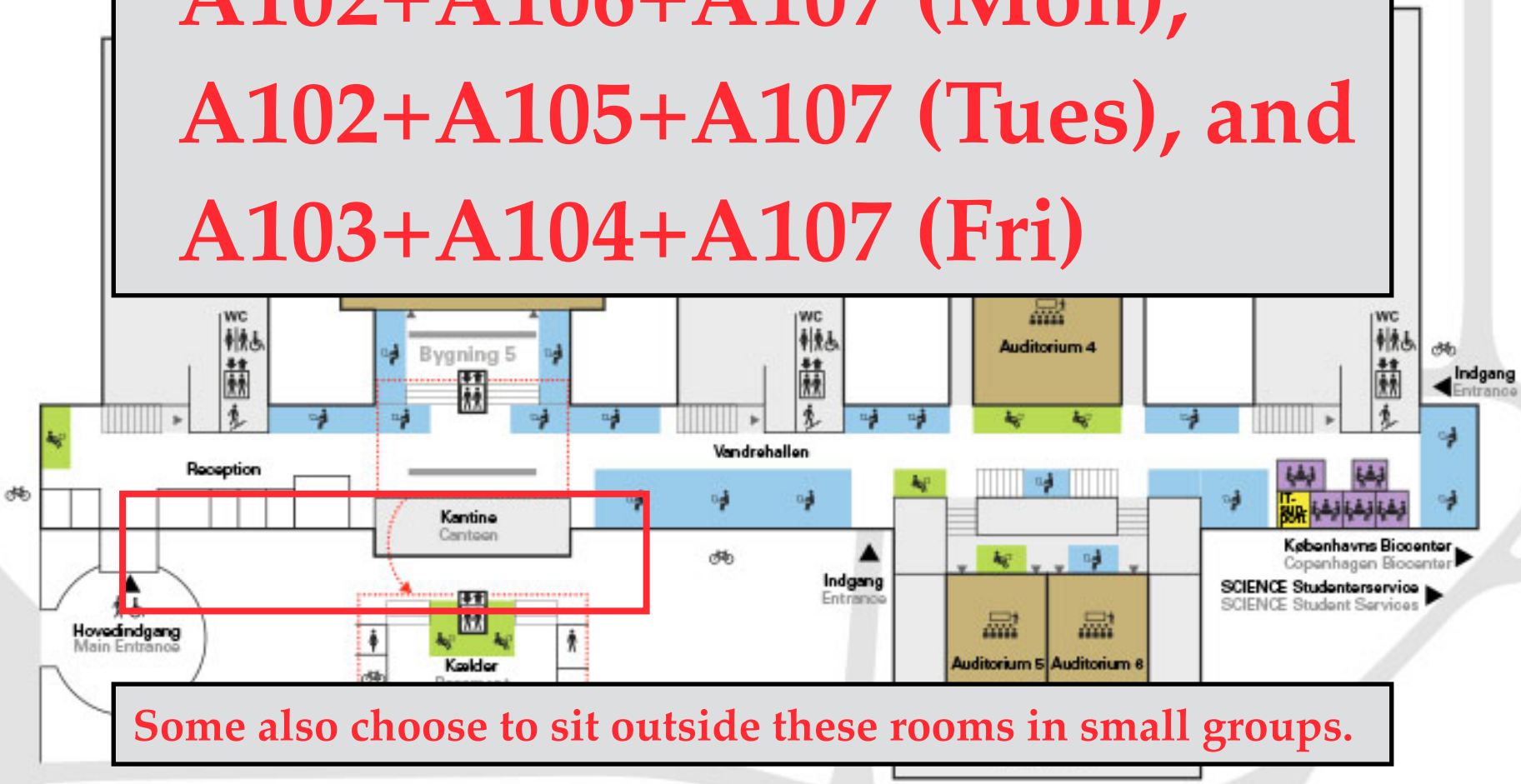
Exercises at HCØ

Exercises:

A102+A106+A107 (Mon),

A102+A105+A107 (Tues), and

A103+A104+A107 (Fri)



Some also choose to sit outside these rooms in small groups.

Additional locations

Hours & Rooms

Hours:

Following block B, but after the first two weeks, we will be using the morning hours 8:15 - 9:00 Monday and Friday for "self-studying", *except for project experiments.*

Monday:

8:15 - 10:00 Lectures
10:15 - 12:00 Exercises

Tuesday:

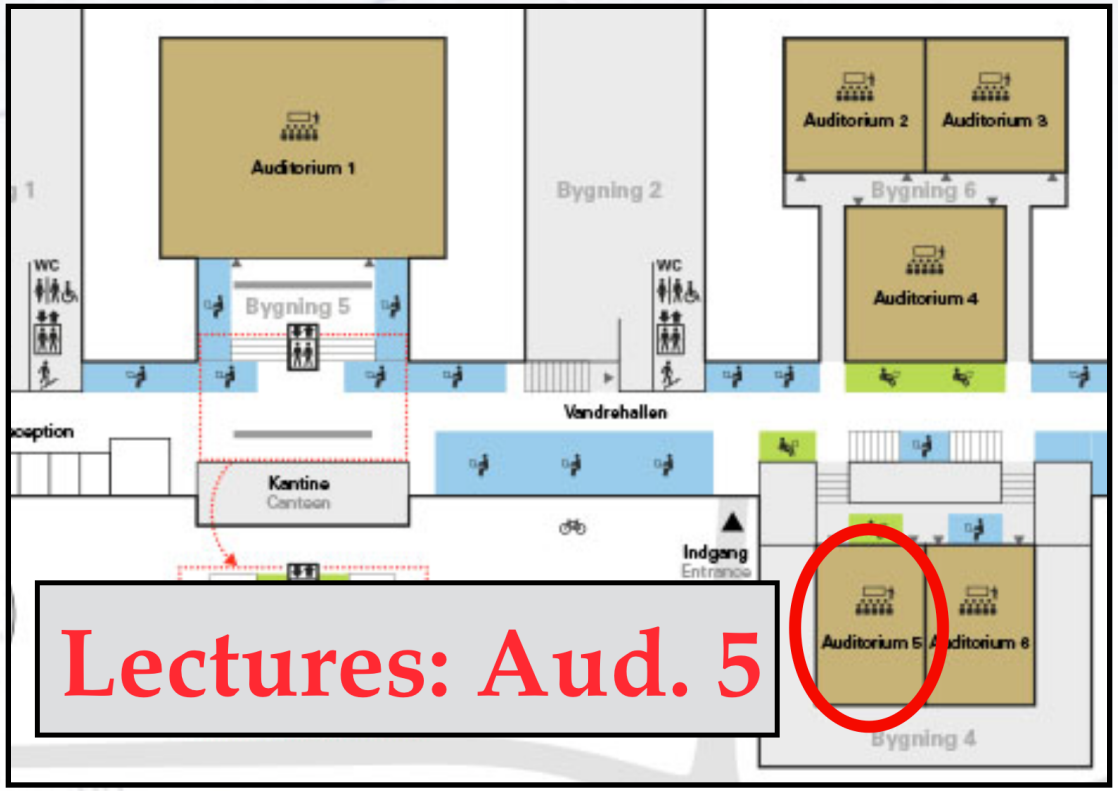
13:15 - 14:00 Lectures
14:15 - 17:00 Exercises

Friday:

8:15 - 10:00 Lectures
10:15 - 12:00 Exercises

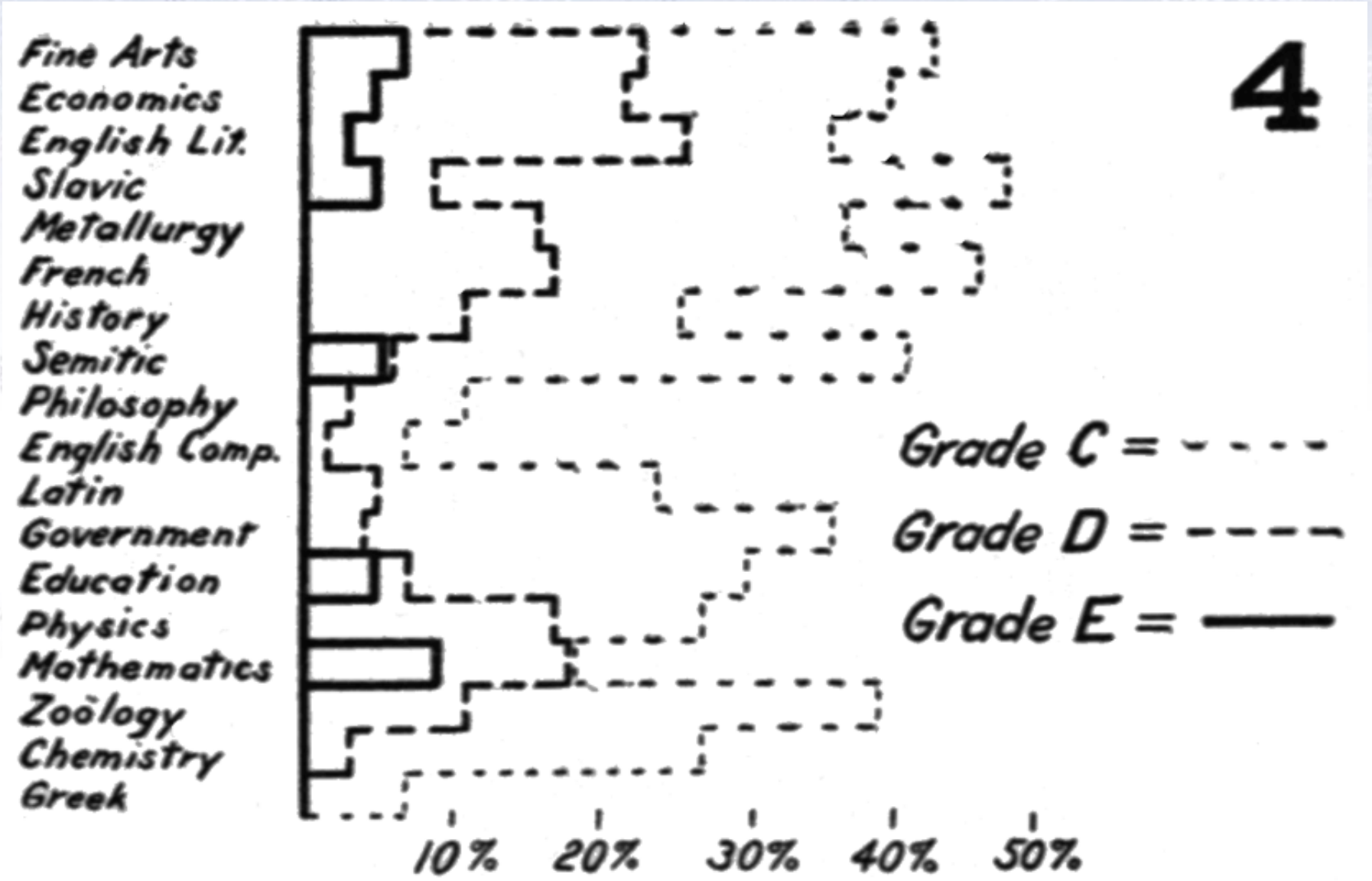
Rooms:

Lectures: HCØ Auditorium 5
Exercises: HCØ Svalegangen



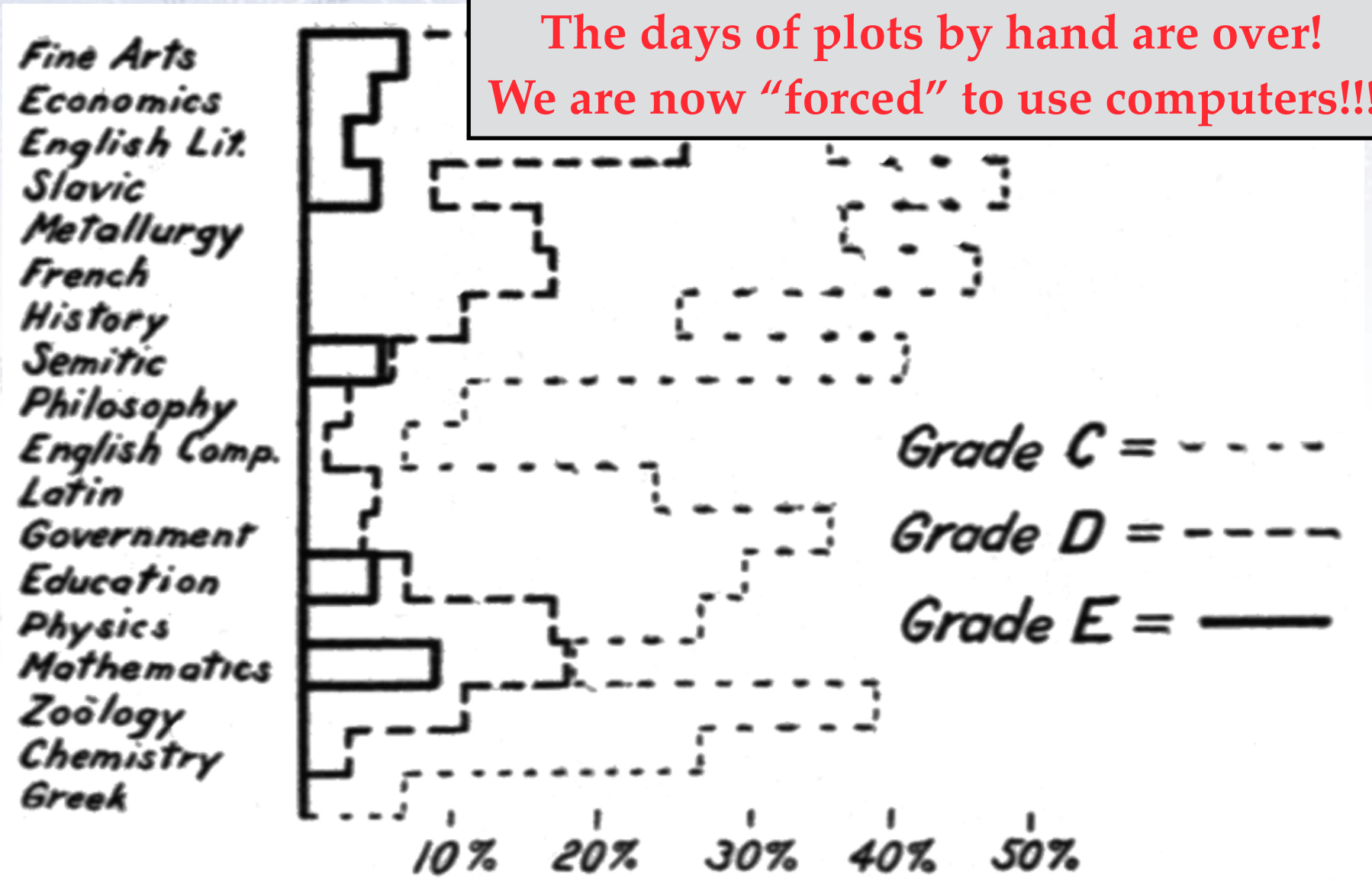
First week: Additional Python introduction Friday 12:15-13:00 in BioCenter!

Computers and software



Computers and software

The days of plots by hand are over!
We are now "forced" to use computers!!!



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**
Programming: **Python** - version 3.6.X
Editor: **Jupyter Notebook** (or own favorit!)

Python Packages used:

- NumPy, MatPlotLib, iMinuit, SciPy, SeaBorn, os, sys, and a few others.
  Only iMinuit should possibly be "unknown" to many, but it is easy to install, and essential for fitting.

Before course start ("Week 0"), we will introduce python, ERDA, github, etc.:
Thursday 7th 15:15-17:00 in Aud. A: Introduction I (mostly setup).
Thursday 15th 10:15-12:00 in Aud. A: Introduction II (mostly programming).

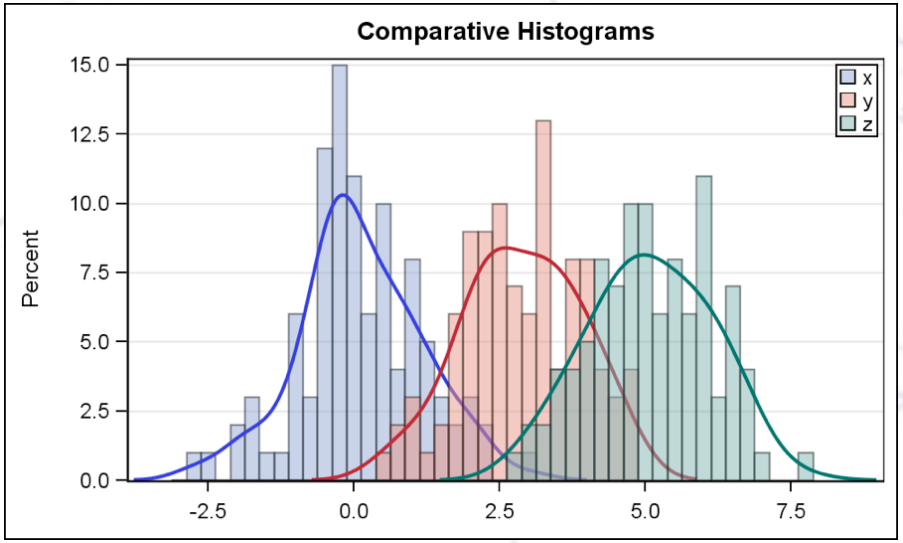
Also, the first Friday after class (i.e. 12:15-13:00), we will further introduce and answer questions on all technical aspects of the course.

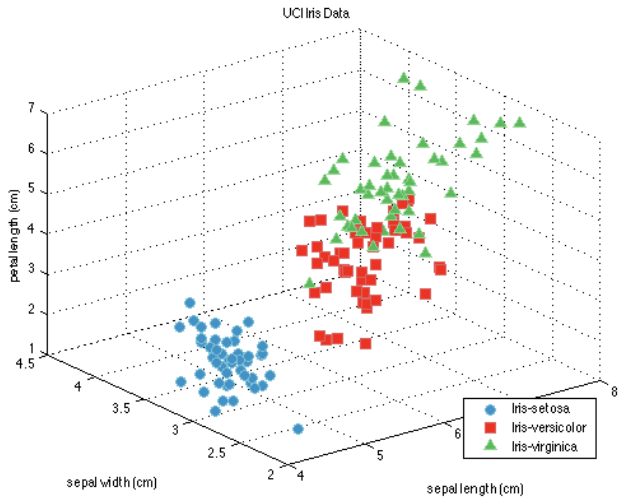
Data sets

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 fields providing ***billions of measurements***.





Literature

We use Roger J. Barlow's "Statistics", as it is an accessible introduction to statistics with many examples, and the best overall book (I think).

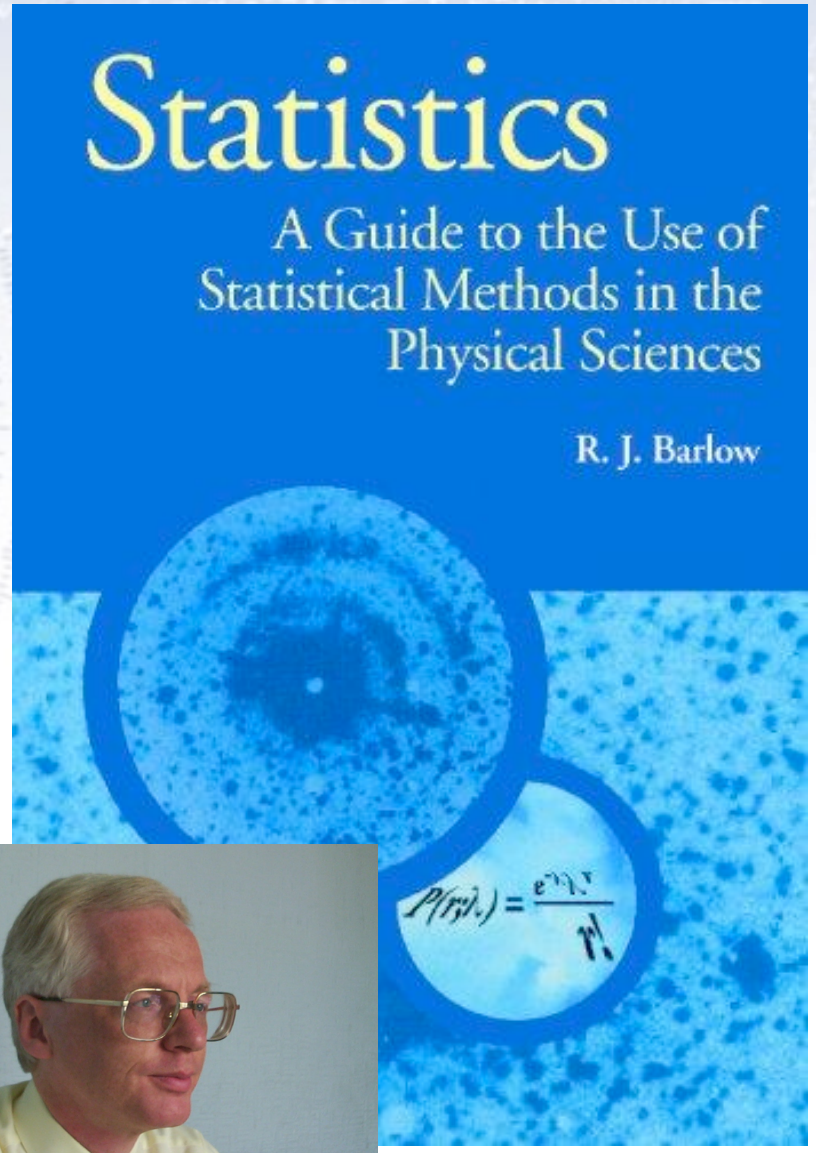
If anything, it is lacking a bit on how to generate random numbers according to a specific PDF and on categorising events.

I might occasionally also refer to:

- Bevington: Data Reduction & Error Analysis
- Cowan: Introduction to Statistics

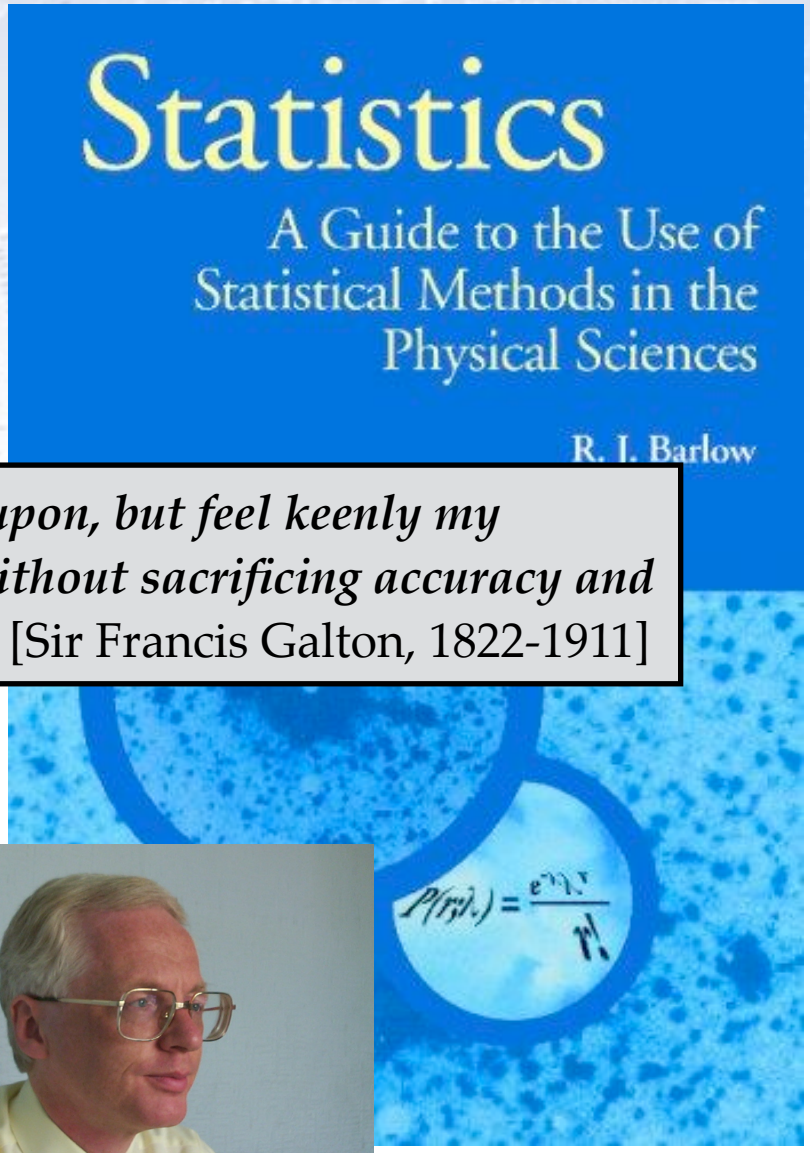
...and notes from Particle Data Group!

NOTE: There is a great abundance of notes, fora, etc. on both statistics but especially also Python on the web, which I encourage you to use (with a proper critical mind).



Literature

We use Roger J. Barlow's "Statistics", as it is an accessible introduction to statistics with many examples, and the best overall book (I think).



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

> ***"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 mi

- B
- Cowan: Introduction to Statistics

...and notes from Particle Data Group!

NOTE: There is a great abundance of notes, fora, etc. on both statistics but especially also Python on the web, which I encourage you to use (with a proper critical mind).

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 most of 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 make sure that you pass the following check list… ***before first day is over***:

- **Have read the course information** (slides on course webpage).
  Otherwise, you don't know what is going to happen.
- **Have your picture ("mug shot") taken** (done in class Monday)**.**
  Otherwise, we don't know who you are.
- **Have filled in the questionnaire** (on course webpage).
  Otherwise, we don't know what you know and don't know.
- **Have measured the length of the lecture table in Auditorium A.**
  Otherwise, you haven't contributed to a common course dataset.
- **Be able to run Python on ERDA and possibly have Python 3.6+ (and a few extra packages) running on your laptop**.
  Otherwise, you can't follow the exercises or solve problems.

In order not to continuously be doing the above, we will finish all of these steps Monday the 18th of November 2019 and (hopefully) not thereafter!

Project

In the second week of the course you will be working on the data analysis following two (simple?) experiments for about two weeks.

They will be in **First Lab** on (dividing class into two halves, the other half having lectures and exercises as normally, TBC):

- Friday the 29th of November 8:15-12:00.
- Monday the 2nd of December 8:15-12:00.

This is your chance to fully do the statistics behind an experiment and play with real data to gain experience of what planning an experiment and detailed data analysis requires! This ***will count 20% in your final grade!!!***

It will require the use of computers and modifications of some of the code you have been running.
You will be working in groups of 4-5 persons, and only one report (2-4 pages) is required from each group.

Real life problems/experiments will resemble this project!

Project

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

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

- Simple pendulum.
- Ball rolling down an incline.

The goal is to **determine g in two ways and propagate the uncertainties** on these measurements. More on that (in time) on the webpages under "project".

Project deadline: One report (in PRL style) per group only is to be handed in by **Sunday the 15th of December 22:00**.

Problem set

During the course, I will give a larger problem set to be solved and handed in. It is due on Sunday the 5th of January by 22:00.

This will cover most of the curriculum covered at this point, and it *will count 20% 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 (even encouraged) to work in groups, but **each student must hand in their own solution**, and you should **state your collaboration**.

The problem set is extensive, so I suggest that you start early.

The final exam will somewhat resemble this problem set!

Exam

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

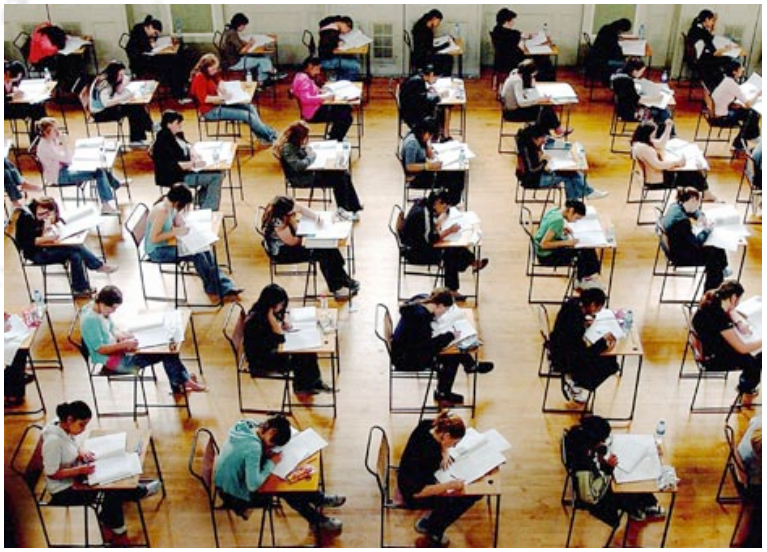
It will cover most of the curriculum, and it ***will count 60% in your final grade!!!***

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 16th of January 8:00am.

It will then naturally have to be handed in:
Friday the 17th of January before 18:00!



(*) 28 "awake" hours, i.e. 8-24 the 16th + 6-18 the 17th.

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 also through your active participation that you have this privilege (i.e. that I'll listen most).

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).

Problems?

If you experience problems in relation to Applied Statistics, whatever their origin and nature, then write me!

I may not be able to do anything about it, but if I don't know about your problems, then I most certainly can not do anything about them.

I consider myself fairly large, as long as I feel that this largeness is met by sincerity and will.

But… you need to write me in the first place! That is your responsibility.

Power/wifi for computers

There are no certainty of individual power sockets in the auditorium nor in the exercise class rooms.

We will try our best to bring a few extension cords, but please charge your laptops before class.

We hope that the Wifi will be good enough in all the rooms. However, it is always good to have a local version of Python running, should this fail.

Course book and References

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

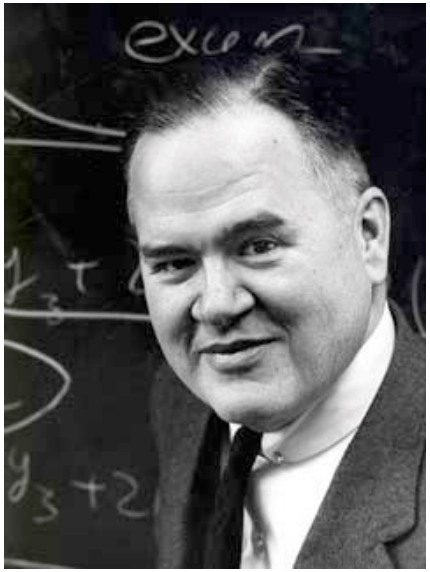
Philip R. Bevington: Data reduction and error analysis.
Classic introduction with very good examples - a standard reference in all of experimental physics.

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, written in a useful or precise way.

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

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