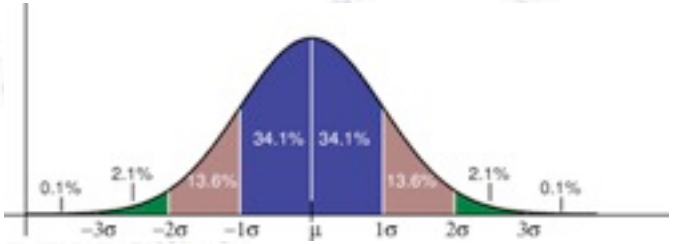
Applied Statistics

Course information






Troels C. Petersen (NBI)



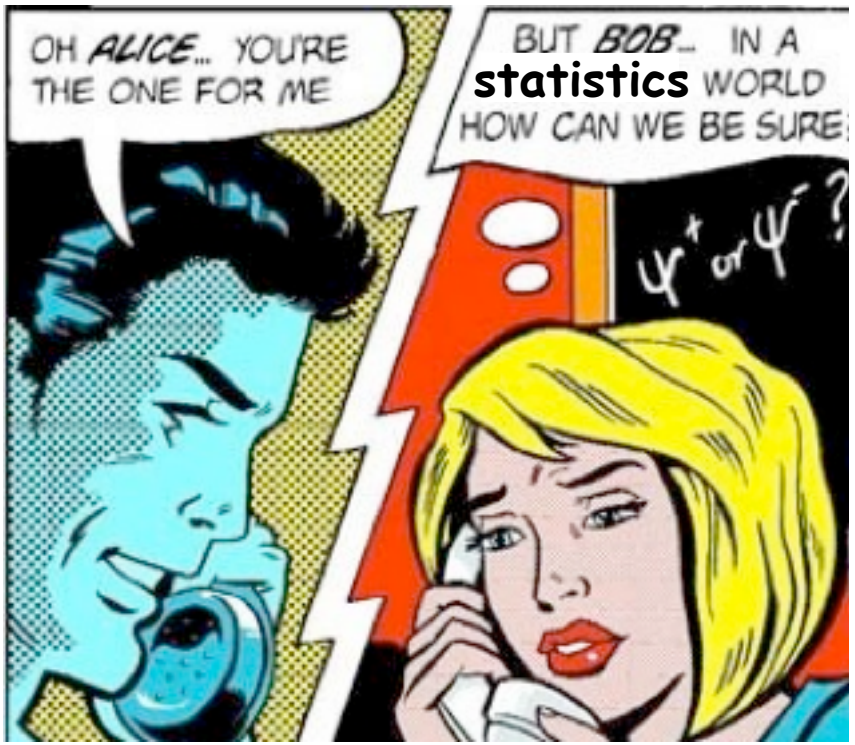
"Statistics is merely a quantization of common sense"

Applied Statistics

All the technical stuff!

Technicals:

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



http://www.nbi.dk/~petersen/Teaching/AppliedStatistics2012.html

Rooms and hours

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

Auditorium M is in building M. Unless otherwise stated, this is it! C++/ROOT intro first 4 weeks.

Monday:
9:15 - 10:00 Lectures
10:15 - 12:00 Exercises

Tuesday:
13:15 - 14:00 Lectures
14:15 - 16:00 Exercises
16:15 - 17:00 C++/ROOT intro

Friday:
9:15 - 10:00 Lectures
10:15 - 12:00 Exercises

Computers and software

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

Operating system: **Linux/MAC OS (or Windows)**
Editor: **Emacs** (or your own favorite!)
Programming: **C++**
Higher level analysis program: **ROOT** (based on C++)

- I've prepared a two-page "cheat sheet" on these.
- Sascha and I will give an introduction today and the last hour of the Tuesday block (in Auditorium M).
  see also: http://www.nbi.dk/~petersen/Teaching/Stat2012/rootintro.html

Alternatively, people can draw their histograms in hand!!!

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!

I've tried my best to search for a large variety of data sets, but this is not easy.

As a result, many 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***.

Alternatively, people will be counting cars out on Blegdamsvej!!!

Literature

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

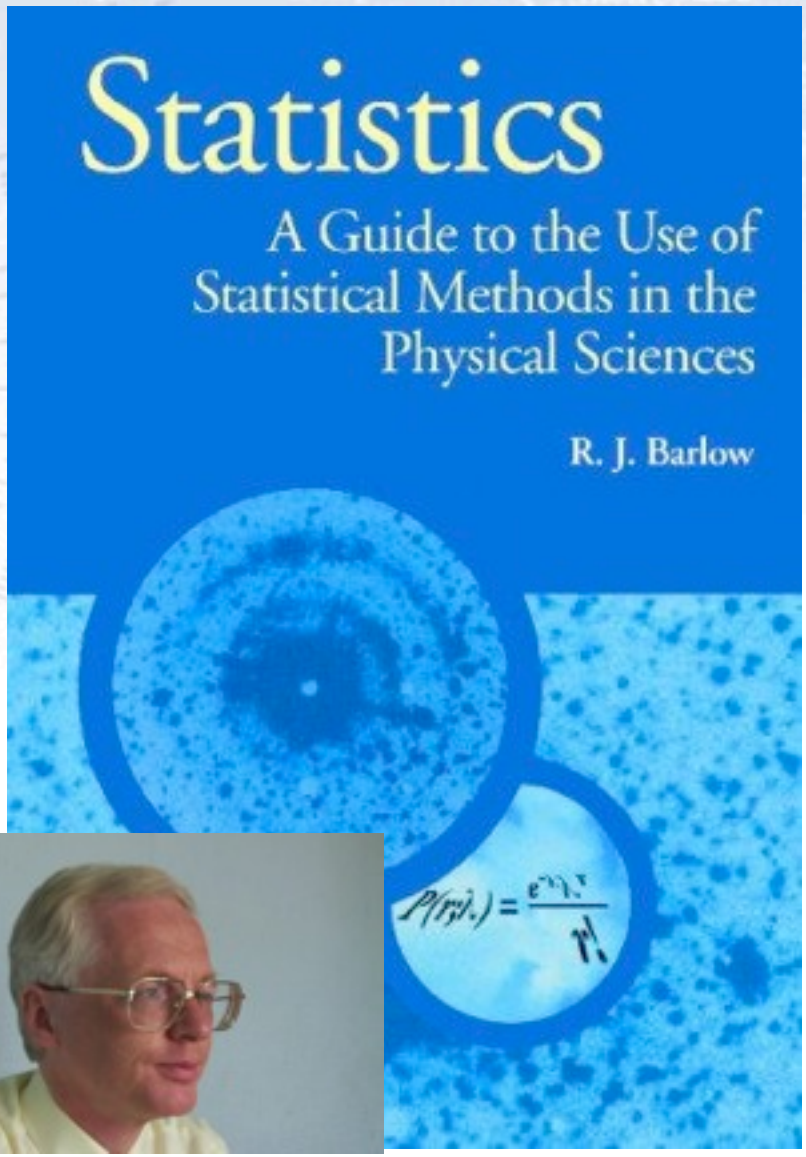
It is a very good and accessible introduction to statistics, and it gives many examples.

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!



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 in this process that statistics is really understood.

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 a detailed data analysis requires!
Each of these ***will count 15% 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 (2-3 pages) is required from each group.

Real life problems will resemble these projects very much!

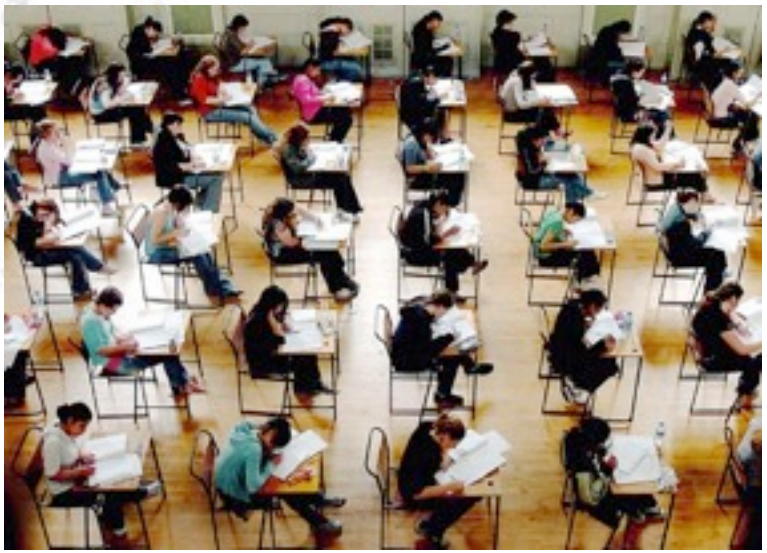
Exam

Exam will be a 24 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 55% in your final grade!!!*

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

You should work on your own!



I will provide this exam on:
Thursday the 1st of November 8:00am.

It will then naturally have to be handed in:
Friday the 2nd of November (before noon!)

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/criticize/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

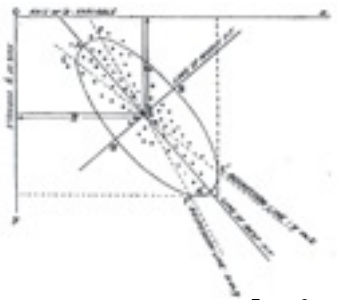
The famous statistician John Tukey (1915-2000) was quoted for wanting to teach:

- 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 analyzed.
- 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.

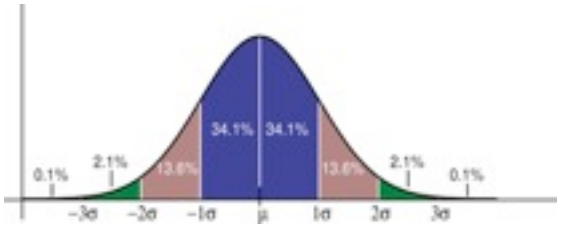
"Far better an approximate answer to the right question, which is often vague, than an exact answer to the wrong question, which can always be made precise." J. W. Tukey

Notes on the ChiSquare method

"It was formerly the custom, and is still so in works on the theory of observations, to derive the method of least squares from certain theoretical considerations, the assumed normality of the errors of the observations being one such.
It is however, more than doubtful whether the conditions for the theoretical validity of the method are realized in statistical practice, and the student would do well to regard the method as recommended chiefly by its comparative simplicity and by the fact that it has ***stood the test of experience****".*
[G.U. Yule and M.G. Kendall 1958]



Top 10



most important things in applied statistics

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