TMS150/MSG400

Stochastic data processing and simulation

Course Aims

- Solve concrete statistical problems using mathematical and statistical software
  - R, Matlab and C
- Communicate your solution to others in writing

Learning objectives

- Apply theoretical knowledge of mathematical statistics in realistic problems
- Combine analytical (pen and paper) with numerical (computer) problem solving
- Practice scientific and technical writing
- Learn/practice software packages: Matlab, R, C, Latex

Course topics

- Work with and explore data sets
- Simulation of data from a given distribution
- Test of distribution assumptions and robust estimators
- Decision theory
- Analysis of financial data, utility maximization
- Bootstrap, empirical distributions, resampling
- Monte Carlo methods, Monte Carlo Integration
- Stochastic processes (simulation)
- Reliability and survival

Course web page

http://www.math.chalmers.se/Stat/Grundutb/CTH/tms150/1718/

Course setup

- No written exam
- 6 mandatory computer projects (labs)
- OK to work in pairs, but each student has to write his/her own report
- OK to work from home, but you will only get help during exercise sessions
- Lab 1 and 5, only answers/ show the exercise teacher. Only pass/non-pass
- Lab 2-4 and 6 – Complete reports
- Max 10 pages per report including figures, but excluding appendix
- The report should be written in Latex.
- Include well-commented code in appendix

Computer room sessions

- 4 hours, Mondays and Thursdays
- Please save all interactions with teachers to computer sessions or lectures
  - Many students, and many questions
- If you really need to come to my office, come on Fridays 13:15-14:00
- Mail is OK for questions with short answers

Exercise teacher

Andreas Petersson

- Helps you in the computer rooms
- Grades labs 4 and 6

Computer exercises

Lab	Topic	Programming language	Examination
Lab 1	Robustness and distribution assumptions	Matlab and R	Only answers
Lab 2	Decision theory	Matlab	Complete report
Lab 3	Reliability and survival	R	Complete report
Lab 4	Bootstrap	R	Complete report
Lab 5	Monte Carlo integration	C and Matlab	Only answers
Lab 6	Simulation of stochastic processes	R	Complete report

Grading

Project	Total	Pass
Lab 1	Pass	Pass
Lab 2	13	7
Lab 3	11	5
Lab 4	14	6
Lab 5	Pass	Pass
Lab 6	10	4
Total	**48**	**22**

- Pass on all labs

Chalmers:
3: Pass on all 6 labs (and 22 points)
4: Pass on all 6 labs and 32 points
5: Pass on all 6 labs and 42 points

GU:
G: Pass on all 6 labs (and 22 points)
VG: Pass on all 6 labs and 37 points

Deadlines

Exercise	Type	Recommended deadline	Final deadline
Lab 1 - Robustness and distribution assumptions	Only answers	Monday 11 Sep, 11.45	Monday 30 Oct, 17.00
Lab 2 - Decision theory	Complete report	Monday 25 Sep, 8.15	Monday 30 Oct, 17.00
Lab 3 - Reliability and survival	Complete report	Monday 2 Oct, 8.15	Monday 30 Oct, 17.00
Lab 4 - Bootstrap	Complete report	Monday 9 Oct, 8.15	Monday 30 Oct, 17.00
Lab 5 - Monte Carlo integration	Only answers	Thursday 12 Oct 11.45	Monday 30 Oct, 17.00
Lab 6 - Simulation of stochastic processes.	Complete report	Monday 23 Oct 8.15	Monday 30 Oct, 17.00

Hand-in of reports

- First hand-in:
  - Send email to statdata.chalmers@analys.urkund.se
- Opportunity to hand in one return per report to retrieve more points
- Second hand-in:
  - Send from same mail address
  - Mention in mail or highlight in report what has changed

Programming

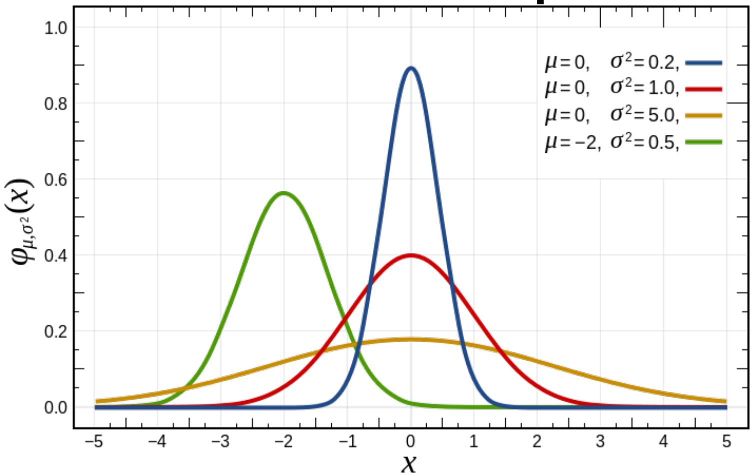
- Matlab
  - Vector operations and matrix multiplications typically faster than for-loops
- R
  - Vector operations and matrix multiplications typically faster than for-loops
  - Statistical programming language
  - Free
  - Load packages/libraries
- C
  - Fast (especially compared to Matlab/R code with many for-loops)
  - Low-level programming language

Programming tips

- Use in-built functions if available
- Check help pages to get details about a function
  - Matlab: `help std`
  - R: `?sd, help('sd')`

Distribution assumptions in statistics

- Model the randomness/variation in the data
- Assumption: A random sample comes from an underlying distribution
- Statistical testing: The test statistic follows a specific distribution under the null hypothesis
- Are the distribution assumptions correct?



Example: Body temperature

- Measurements for 65 male and 65 female patients

Body temperature	Gender	Heart rate
35.7	male	70
36.3	male	69
36.3	male	78
36.4	male	69
36.6	male	73
36.6	male	72
36.7	male	67
36.7	male	67
36.8	male	72
36.9	male	70

Body temperature	Gender	Heart rate
36.9	male	71
37.0	male	66
37.1	male	78
37.2	male	80
37.3	male	71
37.5	male	75
36.2	female	66
36.5	female	84
36.6	female	71
36.7	female	76

Body temperature	Gender	Heart rate
36.7	female	89
36.8	female	65
36.8	female	79
36.9	female	81
37.0	female	77
37.1	female	64
37.1	female	70
37.1	female	73
37.2	female	81
37.8	female	78

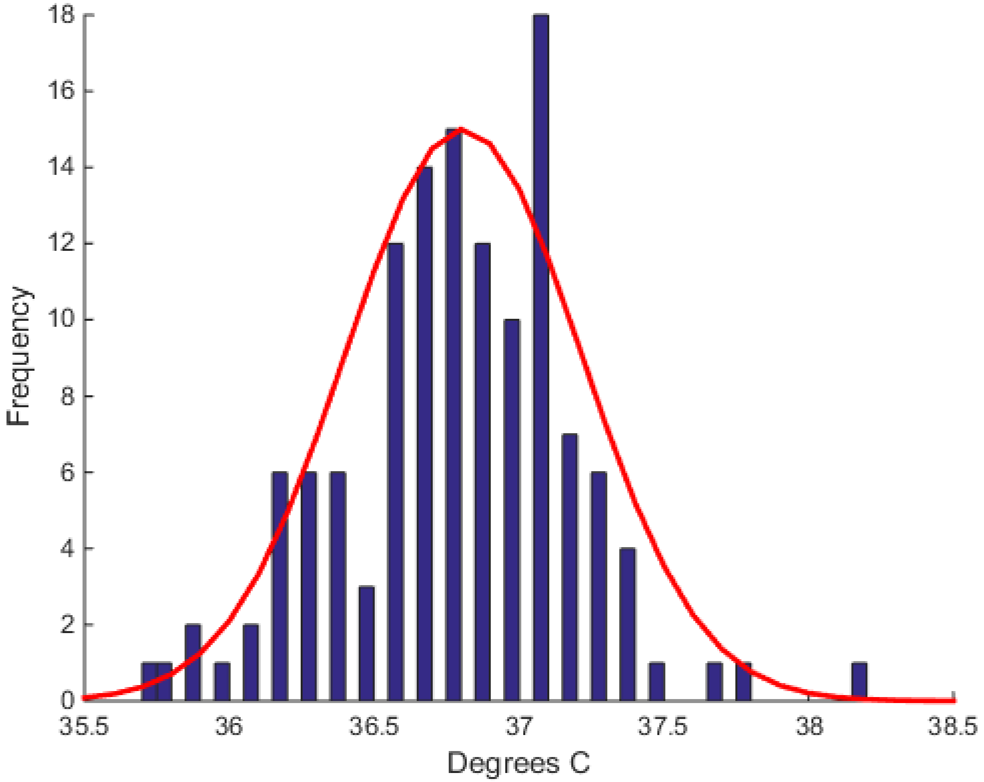
- Mackowiak, P. A., Wasserman, S. S., and Levine, M. M. (1992), "A Critical Appraisal of 98.6 Degrees F, the Upper Limit of the Normal Body Temperature, and Other Legacies of Carl Reinhold August Wunderlich," *Journal of the American Medical Association*, 268, 1578-1580.

Questions to ask

- Is the distribution of temperatures normal?
- Is the true population mean really 37.0 degrees C?
- At what temperature should we consider someone's temperature to be "abnormal"?
- Is there a significant difference between males and females in normal temperature?
- Is there a correlation between body temperature and heart rate?
- Were the original temperatures taken on a Centigrade or Fahrenheit scale?

Distribution assumptions

- Explore/plot the data
  - Histogram



Test of distribution assumptions

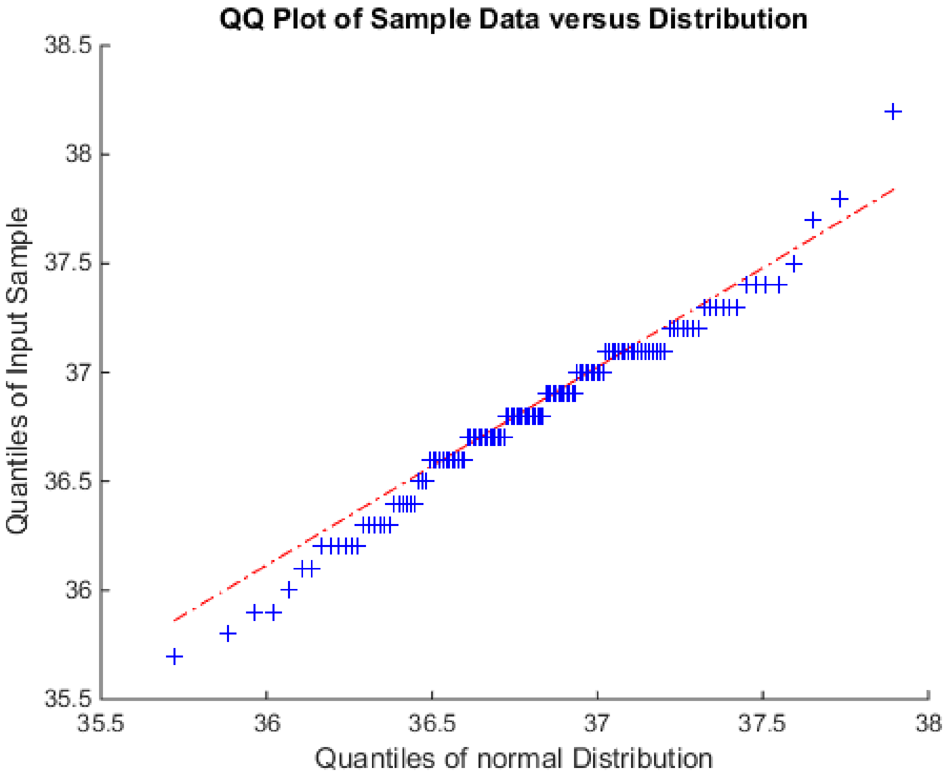
- Graphical
  - Quantile-quantile plot (qq-plot)
  - Probability-probability plot (pp-plot)

- Formal test
  - Chi-square goodness of fit test
  - Kolmogorov-Smirnoff test (KS-test)
  - Other tests (e.g. for specific distributions)

- In both ways we compare the (empirical) distribution of the data with a theoretical distribution

qq-plot (quantile-quantile plot)

- Temperature data



- Assume that the data is normally distributed.
- Estimate the parameters μ and σ^2 from the data
- $\mu=36.808$, $\sigma^2=0.166$
- Compare the distribution of the data with N(36.808,0.166)

If the data is normally distributed the points will approximately follow the line x=y

Graphical test of distribution assumptions

- Investigate if the sample $X_1,\ldots,X_n$ follows a distribution with cdf $F(x)$
- Order the sample: $X_{(1)}, X_{(2)}, \ldots X_{(n)}$
- Plot the sequence of pairs:

qq-plot: $\left\{X_{(i)}, F^{-1}(i/n)\right\}_{i=1}^{n}$

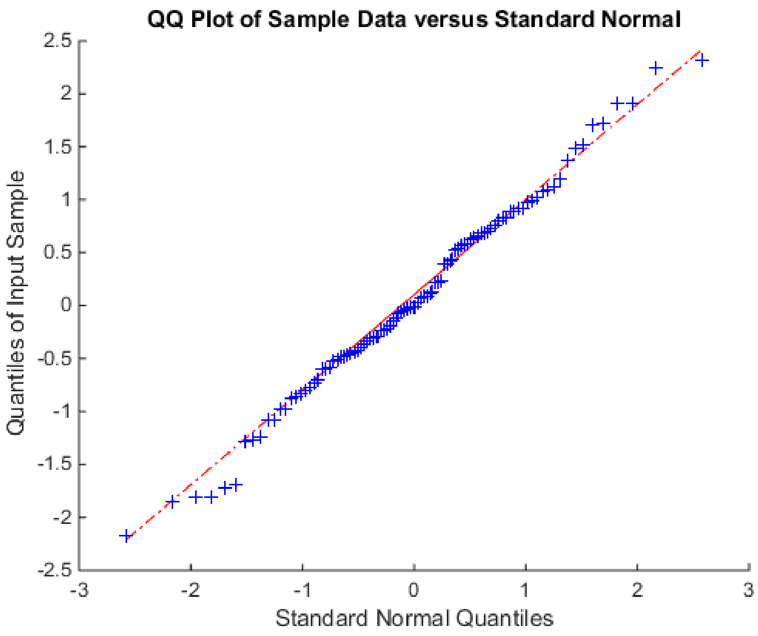
Sometimes *(i-0.5)/n*

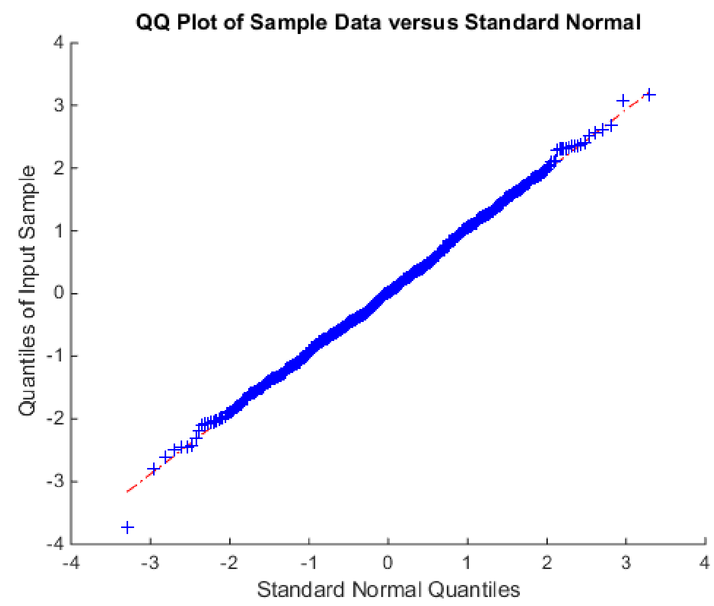
Theoretical quantile

pp-plot: $\left\{\frac{i}{n}, F\left(X_{(i)}\right)\right\}_{i=1}^{n}$

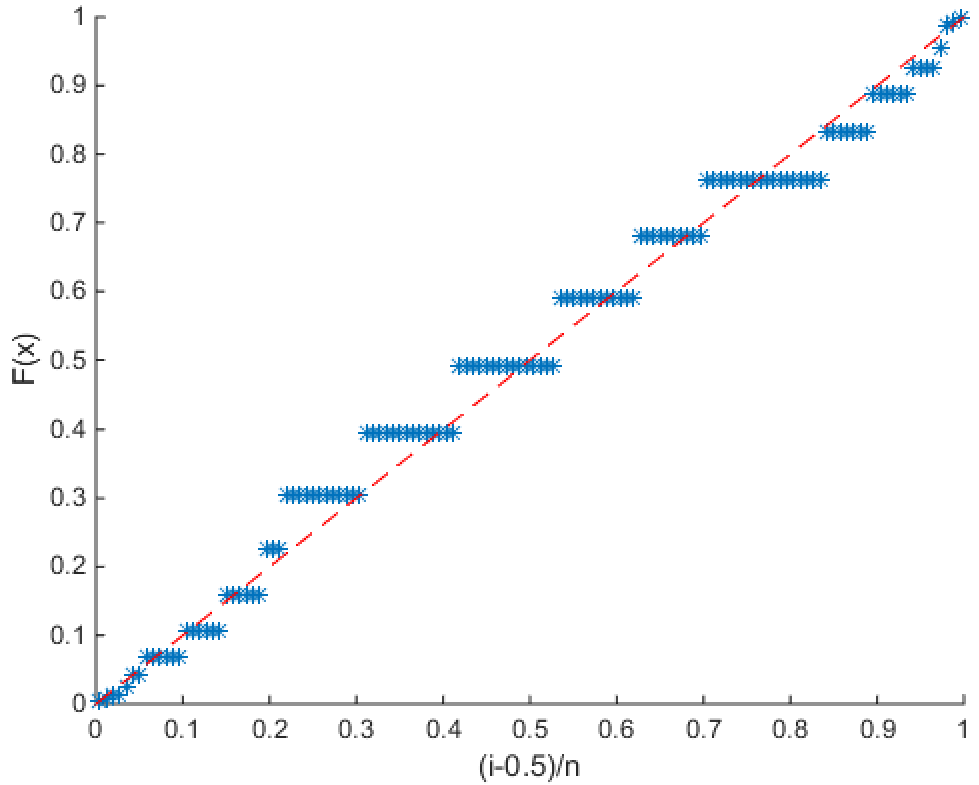
qq-plot (quantile-quantile plot)

- Simulated normal distribution
  - n=100 n=1000





pp-plot



Chi-square Goodness of fit test

- H_0: The data is normally distributed
- H_a: The data is not normally distributed

- Estimate the parameters μ and σ^2 from the data
- Divide the data into k bins

Bin	number of observed	number of expected
35.70-36.20	13	8.83
36.20-36.45	12	15.87
36.45-36.70	29	26.75
36.70-36.95	27	31.30
36.95-37.20	35	25.43
37.20-37.45	10	14.34
37.45-38.20	4	7.47

Expected observations in bin $(a,b]$:
$E_i=(F(b) - F(a))*N$

- Test statistic:

$$\chi^2 = \sum_{i=1}^{k} \frac{(O_i - E_i)^2}{E_i}$$

- Under H_0: Follows approximately a Chi-square distribution with k-1-c degrees of freedom (c is the number of estimated parameters)

Chi-square Goodness of fit test

For the temperature data:

χ^2=10.23

df=7-2-1=4

Critical value: $\chi^2_{0.95,4}$ *=9.48 (one-sided)*

Reject the null hypothesis

p-value=0.037

Kolmogorov Smirnov test (KS-test)

- Measure how much a pp-plot deviates from a 45 degree line (at maximum)
- Test statistic

$$D = \max_{1 \leq i \leq n} \left| F(X_{(i)}; \theta) - \frac{i}{n} \right|$$

- $D\sqrt{n}$ is asymptotically Kolmogorov distributed under the null hypothesis (if the parameters are known)

Kolmogorov Smirnov test (KS-test)

- For the temperature data:
  - KS-statistic: 0.090
  - P-value: 0.232
  - We can not reject the null hypothesis

If the data does not fit the distribution?

- Can the data be transformed in any way?
  - Common transforms: logarithm, square root, square, cube
- Change the model or method
  - Assume another distribution of the data
- Remove outliers
- Robust estimation of parameters

Statistical testing

- Is the mean temperature 37.0 degrees C?
  - H_0: $\mu = 37.0$
  - H_a: $\mu \neq 37.0$

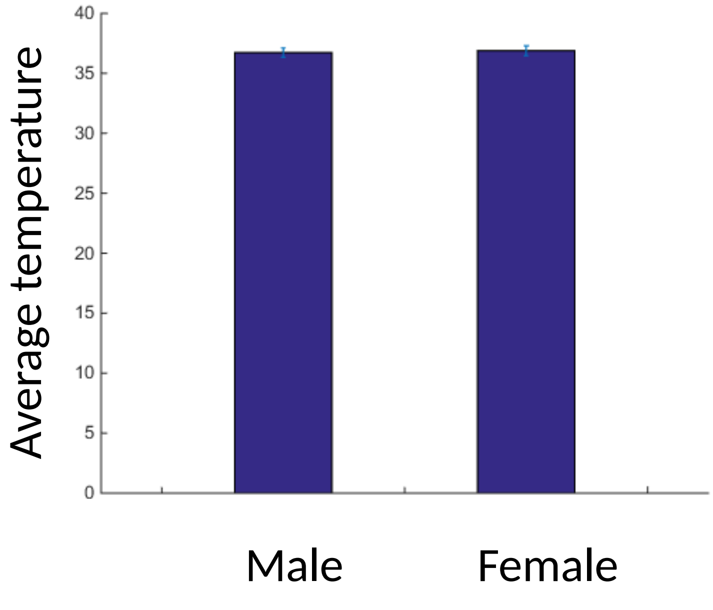
  Matlab: `[h,p,ci,stats]=ttest(temp,37);`

  t-statistic: -5.38, p-value $3.36 * 10^{-7}$

- Does male and female have the same average temperature?
  - H_0: $\mu_1 = \mu_2$
  - H_a: $\mu_1 \neq \mu_2$

  Matlab: `[h,p,ci,stats]=ttest2(temp1,temp2);`

  t-statistic: -2.32, p-value 0.022

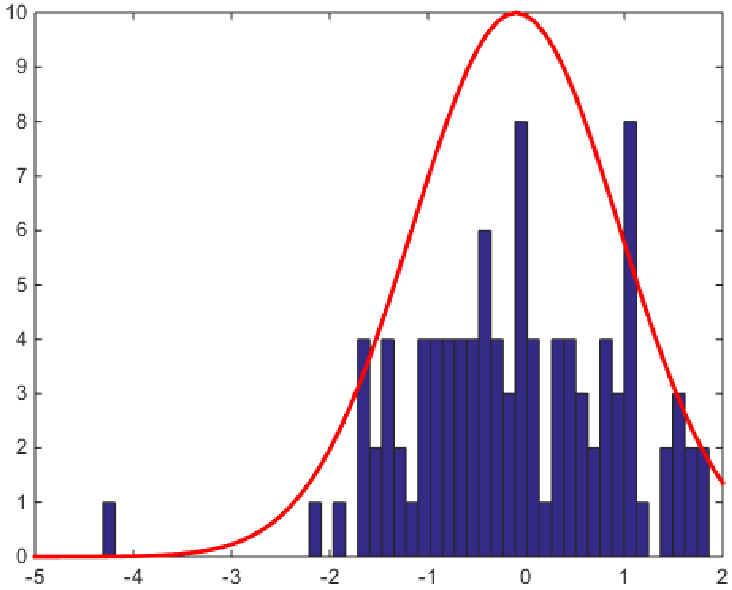


Parameter estimation

- Maximum likelihood estimation
  - Estimate the parameter with the value that maximizes the likelihood of your observations
- Method of moments estimation
  - Express the parameter as a function of distribution moments ($E[X]$, $E[X^2]$, $E[X^3]$, …) and estimate expectations with averages

Robust estimation

- Outliers will affect the parameter estimation
  - Example:



  - $\widehat{\mu_1} = -0.10$
  - $\widehat{\mu_2} = -0.05$ (with the outlier removed)
  - Median, alpha-trimmed mean are robust estimators, but less effective

α-trimmed mean

- Remove outliers
- Sort the observations and take the mean of the ”middle” observations
- *α=0.1*

$$\bar{X}_\alpha = \frac{X_{(k+1)} + \ldots + X_{(n-k)}}{n-2k} \qquad \text{with} \qquad \alpha = \frac{k}{n}$$

Further reading

- Read about qq-plots etc. in a Statistics textbook,
- e.g. chapter 10 in
  Rice, John A. Mathematical Statistics and Data Analysis Third Edition 2007, Brooks/Cole, CENGAGE Learning. ISBN-13-0-495-11868-8

Computer exercise

- Matlab
- R
- No report writing for lab 1
- Pass by showing you answers to Andreas or me during lab sessions or mail them to kallus@chalmers.se