Computational Statistics
A Proposal for a Basic Course

Statistical Computing 2009
28.6.-1.7.2009, Schloss Reisensburg

Günther Sawitzki
<gs@statlab.uni-heidelberg.de>

StatLab Heidelberg

30. 6.2009
Typeset: July 26, 2009

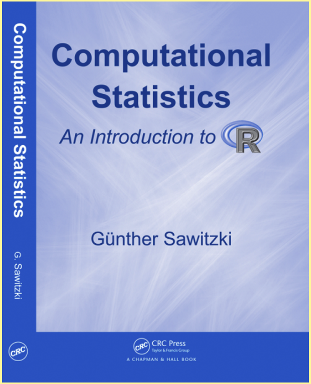
Note

Page references refer to

Computational Statistics:
An Introduction to R

Chapman & Hall/CRC Press,
Boca Raton (FL), 2009.

ISBN: 978-1-4200-8678-2



See http://sintro.r-forge.r-project.org/.

Background

Background

Predecessors

Aim

A concise course in computational statistics

Predecessors

- One week post-graduate course "Biometry in Medicine"
- One week course: R programming
- Linear Models
- Statistical Data Analysis
- ...

Background

Predecessors

Aim

A concise course in computational statistics

Predecessors

- One week post-graduate course "Biometry in Medicine"
- One week course: R programming
- Linear Models
- Statistical Data Analysis
- ...

Background

Predecessors

Aim

A concise course in computational statistics

Predecessors

- One week post-graduate course "Biometry in Medicine"
- One week course: R programming
- Linear Models
- Statistical Data Analysis
- . . .

Background

Predecessors

Aim

A concise course in computational statistics

Predecessors

- One week post-graduate course "Biometry in Medicine"
- One week course: R programming
- Linear Models
- Statistical Data Analysis
- ...

Background

Predecessors

Aim

A concise course in computational statistics

Predecessors

- One week post-graduate course "Biometry in Medicine"
- One week course: R programming
- Linear Models
- Statistical Data Analysis
- ...

Background

Audience

Computational Statistics: An Introduction to R

Designed for a mixed audience

- researchers and post-graduates from applied areas (in particular from clinical departments and from the DKFZ, the German cancer research center), with some working knowledge in statistical methods and with considerable laboratory experience
- students from mathematics or computer science, with a basic knowledge in (mathematical) stochastics

As one of the participants from the applied field said "We can look up the methods ourselves. What we need is a guide to the underlying concepts."

Background

Audience

Computational Statistics: An Introduction to R

Designed for a mixed audience

- researchers and post-graduates from applied areas (in particular from clinical departments and from the DKFZ, the German cancer research center), with some working knowledge in statistical methods and with considerable laboratory experience
- students from mathematics or computer science, with a basic knowledge in (mathematical) stochastics

As one of the participants from the applied field said ”We can look up the methods ourselves. What we need is a guide to the underlying concepts.”

Background

Topics

What do we need?

Try to illustrate/demonstrate:

What are the statistical concepts and methods that are essential for computational statistics on a scientific level?

What is not needed?

How to survive bolognese?

Background

Topics

What do we need?

Try to illustrate/demonstrate:

What are the statistical concepts and methods that are essential for computational statistics on a scientific level?

What is not needed?

How to survive bolognese?

Background

Topics

What do we need?

Try to illustrate/demonstrate:

What are the statistical concepts and methods that are essential for computational statistics on a scientific level?

What is not needed?

How to survive bolognese?

Background

Topics

Statistical Topics

Idea: Select a small set of fairly general statistical topics.

Background

Topics

Statistical Topics

- distribution diagnostics

  given $X_i \qquad i = 1, \ldots, n, \qquad$ infer on $\mathscr{L}(X_i)$
- regression models and regression diagnostics

  $Y = m(X) + \varepsilon$
- non-parametric comparisons
- multivariate analysis

Background

Topics

Statistical Topics

- distribution diagnostics

  $$\text{given } X_i \qquad i = 1, \ldots, n, \qquad \text{infer on } \mathscr{L}(X_i)$$

- regression models and regression diagnostics

  $$Y = m(X) + \varepsilon$$

- non-parametric comparisons
- multivariate analysis

Background

Topics

Statistical Topics

- distribution diagnostics

  given $X_i \qquad i = 1, \ldots, n, \qquad \text{infer on } \mathscr{L}(X_i)$
- regression models and regression diagnostics

  $Y = m(X) + \varepsilon$
- non-parametric comparisons
- multivariate analysis

Background

Topics

Statistical Topics

- distribution diagnostics

  given $X_i \qquad i = 1, \ldots, n, \qquad$ infer on $\mathscr{L}(X_i)$
- regression models and regression diagnostics

  $Y = m(X) + \varepsilon$
- non-parametric comparisons
- multivariate analysis

Background

Topics

Note

Topics refer to statistical problem classes,

not specifically to heuristics such as least square, maximum likelihood etc.,

not to specific models.

They try to mark a broad range of topics.

Topics may be used as self-contained teaching modules, with only limited cross-import. They can be taught as separate units.

Background

Topics

The course may be presented as an introduction to R. But actually it is an invitation to statistical data analysis.

Time Table

Compact course (5 days).

or

One term, 2h lectures plus 2h exercises per week.

Details to come.

Background

Structure

Chapter Structure

- Content chapters - used as course material.
  - Core content
  - R supplement
  - Statistical summary

Course Material Structure

- Four chapters, by statistical topic - used as course material.
- Appendix: R Reference sections by programming topic - used as supplement or for look up

Contents

Contents

Ch. 2: Distribution Diagnostics

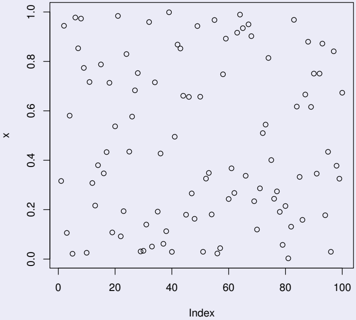
Contents

Ch. 2: Distribution Diagnostics

Example 1.1: A Simple Plot (p.7)

Input

```
x <- runif(100)
plot(x)
```



Contents

Ch. 2: Distribution Diagnostics

Exercise 1.1(p.7)

Try experimenting with these plots and *runif()*. Do the plots show images of random numbers?
To be more precise: do you accept these plots as images of 100 independent realisations of random numbers, distributed uniformly on $(0, 1)$?
Repeat your experiments and try to note as precisely as possible the arguments you have for or against (uniform) randomness. What is your conclusion?

Walk through your arguments and try to draft a test strategy to analyse a sequence of numbers for (uniform) randomness. Try to formulate your strategy as clearly as possible.

Contents

Ch. 2: Distribution Diagnostics

Exercise 1.2 (p.10)

Use

```
plot(sin(1:100))
```

to generate a plot of a discretised sine function. Use your strategy from Exercise 1.1.

Does your strategy detect that the sine function is not a random sequence?

Contents

Ch. 2: Distribution Diagnostics

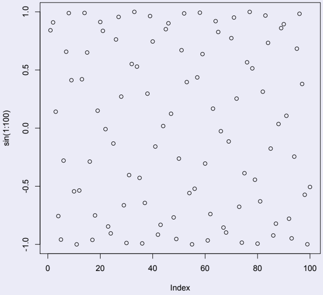
Exercise 1.2 (p.10)

Use

```
plot(sin(1:100))
```

to generate a plot of a discretised sine function. Use your strategy from Exercise 1.1.

Does your strategy detect that the sine function is not a random sequence?



Contents

Ch. 2: Distribution Diagnostics

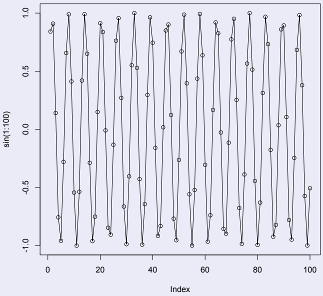
Exercise 1.2 (p.10)

Use

plot(sin(1:100))

to generate a plot of a discretised sine function. Use your strategy from Exercise 1.1.

Does your strategy detect that the sine function is not a random sequence?



Contents

Ch. 2: Distribution Diagnostics

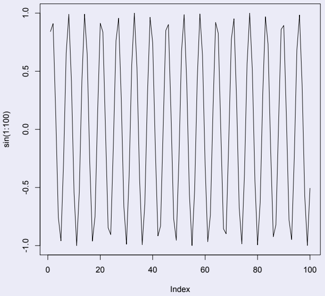
Exercise 1.2 (p.10)

Use

```
plot(sin(1:100))
```

to generate a plot of a discretised sine function. Use your strategy from Exercise 1.1.

Does your strategy detect that the sine function is not a random sequence?



Note

Try to put a challenge.
This is not something to solve on the fly.
It is something to come back to.

Running Exercise in Ch. 1:

Can you tell a uniform from a Gaussian distribution, based on a sample?
Various methods are discussed.

What is the minimum sample size at which the distribution is barely recognizable?

What is the sample size needed for a clear impression?

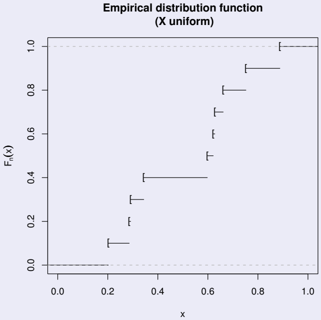
Contents

Ch. 2: Distribution Diagnostics

Case Study: Distribution Diagnostics (p.10)

- Distribution function
- Histogram
- Smoothing (kernel density estimation)

Exercise 1.4 (p.16)



Note

We start with possibly competing approaches. There is more than one way.
For each approach, the theory (and pragmatics) is developed in steps.
After each step, the question is addressed how these approaches compare, and, ultimately, whether there is one which is to be preferred.

Contents

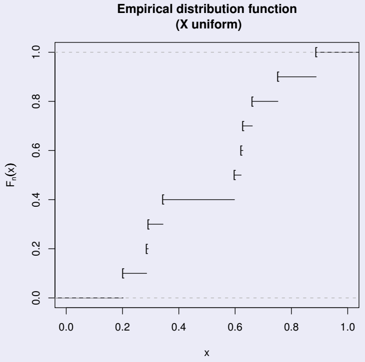
Ch. 2: Distribution Diagnostics

Case Study: Distribution Function (p.10)

- start with simple prototypes
- refine software, e.g. graphics
- make mathematics correct.

  This may need some theorems, e.g.

Theorem: $F(X_{(i)})$ has a beta distribution $\beta(i, n-i+1)$.
Corollary:
$E(F(X_{(i)})) = i/(n+1)$.

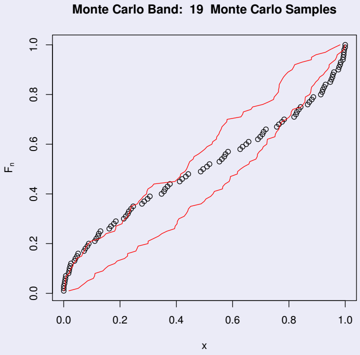


Contents

Ch. 2: Distribution Diagnostics

Example 1.11 Monte Carlo Confidence Bands (p.23)

Simulation can also help us to get an impression of the typical fluctuation. We use random numbers to generate a small number of samples, and compare our sample in question with these simulations. For comparison, we generate envelopes of these simulations and check whether our sample lies within the area delimited by the envelopes.



From: Statistical Computing 1993

We claim: a diagnostic plot is only as good as the hard statistical theory that is supporting it.

G. Sawitzki in:
Computational Statistics. Papers collected on the Occasion of the 25th Conference on Statistical Computing at Schloss Reisensburg.
Edited by P.Dirschedl & R.Ostermann for the Working Groups . . .
Physica/Springer: Heidelberg, 1994, isbn 3-7908-0813-x, p. 237-258.

Plots and their statistical counterparts

Plot	**Statistics/Test**
histogram	χ^2 tests
distribution function	Kolmogorov - Smirnov test

Contents

Ch. 2: Distribution Diagnostics

Statistics for the probability plot (p.27)

Theorem: (Kolmogorov, Smirnov) For a continuous distribution function F, the distribution of $\sup_x |F_n - F|(x)$ is independent of F (in general, it will depend on n).

Theorem: (Kolmogorov) For a continuous distribution function F and $n \to \infty$ the statistic $\sqrt{n} \sup |F_n - F|$ has asymptotically the distribution function $F_{Kolmogorov-Smirnov}(y) = \sum_{m \in \mathbb{Z}} (-1)^m e^{-2m^2y^2}$ for $y > 0$.

Theorem: (Massart 1990) For all integer n and any positive λ, we have $P(\sqrt{n} \sup |F_n - F| > \lambda) \leq 2e^{-2\lambda^2}$.

Note

With all respect: asymptotics should be put in its place.

Learning the difference between asymptotic statements (such as Kolmogorov) and finite sample bounds (like the Dvoretzky - Kiefer - Wolfowitz inequality studied by Massart) should start early.

Contents

Ch. 2: Distribution Diagnostics

Exercise 1.25 Sample Size (p. 41)

Generate a *PP* plot of the $t(\nu)$ distribution against the standard normal distribution in the range $0.01 \leq p \leq 0.99$ for $\nu = 1, 2, 3, \ldots$.

Generate a *QQ* plot of the $t(\nu)$ distribution against the standard normal distribution in the range $-3 \leq x \leq 3$ for $\nu = 1, 2, 3, \ldots$.

How large must ν be so that the t distribution is barely different from the normal distribution in these plots?
How large must ν be so that the t distribution is barely different from the normal distribution if you compare the graphs of the distribution functions?

See also (p. 42 - p. 45).

Note

Where possible, we try to complement theoretical results by simulations.

At this step, we avoid concepts like power. Instead we draw the attention to the question: what is the sample size we need to solve a certain task?

At this early point of the course, power differences are discussed in terms of required sample size.

We avoid to introduce the term “relative efficiency”, not to overload the chapter.

Contents

Ch. 2: Linear Models and Regression Diagnostics

- General Regression Model
- Linear Model
- Variance Decomposition by Orthogonal Complements, and Analysis of Variance
- Simultaneous Inference
- Beyond Linear Regression

Contents

Ch. 2: Linear Models and Regression Diagnostics

Contents

Ch. 2: Linear Models and Regression Diagnostics

- General Regression Model
- Linear Model
- Variance Decomposition by Orthogonal Complements, and Analysis of Variance
- Simultaneous Inference
  - Scheffé's Confidence Bands (see p. 85 ff)
  - Tukey's Confidence Intervals (see p. 87)
  - Case Study: Titre Plates (see p. 88ff)
- Beyond Linear Regression

Contents

Ch. 2: Linear Models and Regression Diagnostics

- General Regression Model
- Linear Model
- Variance Decomposition by Orthogonal Complements, and Analysis of Variance
- Simultaneous Inference
- Beyond Linear Regression

  Just mentioned:
  - Transformations
  - Generalised Linear Models
  - Local Regression

Contents

Ch. 2: Linear Models and Regression Diagnostics

Note

This chapter: mainly textbook material by now.
With some extensions for a data analytical point of view ...

Still needed: point out the special role of the one dimensional response situation, e.g. as expressed by the Gauss-Markov theorem.

Contents

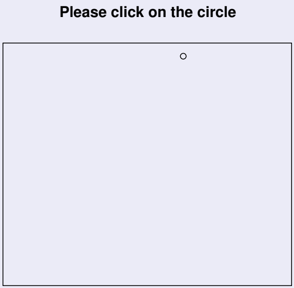
Ch. 3: Non-parametric Comparisons

Contents

Ch. 3: Non-parametric Comparisons

Exercise 3.2: Click Comparison (p. 109)

Try clicking on a random point, with left and then with right hand.



Contents

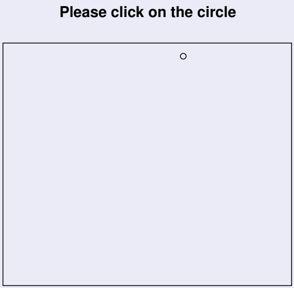
Ch. 3: Non-parametric Comparisons

Exercise 3.2: Click Comparison (p. 109)

Try clicking on a random point, with left and then with right hand.

Immediate impression:
"feels different"

One hand is more responsive.



Contents

Ch. 3: Non-parametric Comparisons

Stochastic Order

Notation: A distribution with distribution function F_1 is *stochastically smaller* than a distribution with distribution function F_2 (in symbols, $F_1 \prec F_2$), if a variable distributed as F_1 takes rather smaller values than a variable distributed as F_2. This means that F_1 increases sooner:
$F_1(x) \geq F_2(x)\ \forall x$ and $F_1(x) > F_2(x)$ for at least one x.

Shift/Scale Families

Notation: For a distribution with distribution function F the family $F_a(x) = F(x - a)$ is called the *shift family* for F. The parameter a is called the shift or location parameter.

Define shift/scale family, and relate to stochastic order.

Note

Stochastic order and stochastic monotonicity are the core concepts that explain why in some situations statistical problems can be reduced to optimization problems.

It is at the core of much of theoretical statistics, e.g. Neyman-Pearson theory.

Recognizing stochastic order relation and stochastic monotonicity are a key competence in statistics.

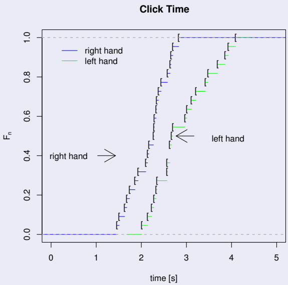
"Monotone likelihood ratios" etc. only obscure the core argument.

Contents

Ch. 3: Non-parametric Comparisons

Two Sample Comparison

Challenge: compare two samples.



Distribution functions for the right/left click time

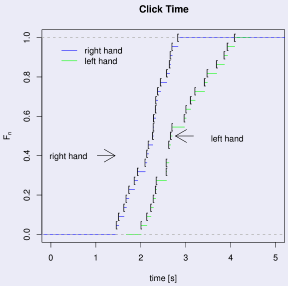
Contents

Ch. 3: Non-parametric Comparisons

Two Sample Comparison

Challenge: compare two samples.

Note: this is not a shift alternative.



Distribution functions for the right/left click time

Contents

Ch. 3: Non-parametric Comparisons

Two Sample Comparison

Challenge: compare two samples.

- $t-$test for normal shift families - recall from Ch. 2
- rank tests (Wilcoxon)
- permutation tests
- bootstrap
- QQ plot
- PP plot & Kolmogorov-Smirnov

Contents

Ch. 3: Non-parametric Comparisons

Two Sample Comparison

Challenge: compare two samples.

- $t-$test for normal shift families - recall from Ch. 2
- rank tests (Wilcoxon)
- permutation tests
- bootstrap
- *QQ* plot
- *PP* plot & Kolmogorov-Smirnov

Contents

Ch. 3: Non-parametric Comparisons

Two Sample Comparison

Challenge: compare two samples.

- $t-$test for normal shift families - recall from Ch. 2
- rank tests (Wilcoxon)
- permutation tests
- bootstrap
- QQ plot
- PP plot & Kolmogorov-Smirnov

Contents

Ch. 3: Non-parametric Comparisons

Two Sample Comparison

Challenge: compare two samples.

- $t-$test for normal shift families - recall from Ch. 2
- rank tests (Wilcoxon)
- permutation tests
- bootstrap
- *QQ* plot
- *PP* plot & Kolmogorov-Smirnov

Contents

Ch. 3: Non-parametric Comparisons

Two Sample Comparison

Challenge: compare two samples.

- $t-$test for normal shift families - recall from Ch. 2
- rank tests (Wilcoxon)
- permutation tests
- bootstrap
- QQ plot
- PP plot & Kolmogorov-Smirnov

Contents

Ch. 3: Non-parametric Comparisons

Two Sample Comparison

Challenge: compare two samples.

- $t-$test for normal shift families - recall from Ch. 2
- rank tests (Wilcoxon)
- permutation tests
- bootstrap
- QQ plot
- PP plot & Kolmogorov-Smirnov

Contents

Ch. 3: Non-parametric Comparisons

Lessons / Issues to point out

Statistical methods / tests may have different assumptions on the data to make them valid.

Statistical methods / tests may have different targets. Which of these methods targets shift alternatives? Which have a more general target?

Unsatisfied assumptions or failed targets do not necessarily imply that a method is not usable. For example, you can use a test targeted at shift alternatives to detect differences which are not covered by shift alternatives.

A thorough discussion is needed here to prepare for the next question: how to compare tests, or methods in general.

Contents

Ch. 3: Non-parametric Comparisons

Two Sample Comparison: Comparison of Methods

Two sample comparisons

- $t-$test for normal shift families - recall from Ch. 2
- rank tests (Wilcoxon)
- permutation tests
- bootstrap
- *QQ* plot
- *PP* plot & Kolmogorov-Smirnov

Contents

Ch. 3: Non-parametric Comparisons

Two Sample Comparison: Comparison of Methods

Two sample comparisons

- $t-$test for normal shift families - recall from Ch. 2
- rank tests (Wilcoxon)
- permutation tests
- bootstrap
- QQ plot
- PP plot & Kolmogorov-Smirnov

Comparison of methods

- sample size comparisons (relative efficiency)
- theoretical power
- power comparison by simulation
- "test beds" and scenarios

Contents

Ch. 3: Non-parametric Comparisons

Two Sample Comparison: Comparison of Methods

Two sample comparisons

- $t-$test for normal shift families - recall from Ch. 2
- rank tests (Wilcoxon)
- permutation tests
- bootstrap
- QQ plot
- PP plot & Kolmogorov-Smirnov

Comparison of methods

- sample size comparisons (relative efficiency)
- theoretical power
- power comparison by simulation
- "test beds" and scenarios

Contents

Ch. 3: Non-parametric Comparisons

Two Sample Comparison: Comparison of Methods

Two sample comparisons

- $t-$test for normal shift families - recall from Ch. 2
- rank tests (Wilcoxon)
- permutation tests
- bootstrap
- QQ plot
- PP plot & Kolmogorov-Smirnov

Comparison of methods

- sample size comparisons (relative efficiency)
- theoretical power
- power comparison by simulation
- "test beds" and scenarios

Contents

Ch. 3: Non-parametric Comparisons

Note

Comparison of methods, other from the sample size point of view, has been postponed until there is a sufficient collection of methods in competition.

There is no discussion of optimality in this course, except for marginal remarks.

"Optimality" is helpful if there is a one dimensional optimality criterion. It may be a misleading focus, if there is more than one aspect to cover.

Contents

Ch. 3: Non-parametric Comparisons

Open Question

What is the state of the art information we should give about two sample comparison, keeping in mind that there are more possibilities for differences than what is covered by shift alternatives?

Contents

Ch. 4: Multivariate Analysis

- Dimensions
- Selections
- Projections
- Sections, Conditional Distributions and Coplots
- Transformations and Dimension Reduction
- Higher Dimensions
- High Dimensions

Contents

Ch. 4: Multivariate Analysis

- Dimensions
- Selections
- Projections
  - Marginal Distributions and Scatter Plot Matrices
  - Projection Pursuit
  - Projections for Dimensions 1, 2, 3, . . . 7
  - Parallel Coordinates
- Sections, Conditional Distributions and Coplots
- Transformations and Dimension Reduction
- Higher Dimensions
- High Dimensions

Contents

Ch. 4: Multivariate Analysis

Contents

Ch. 4: Multivariate Analysis

Open Questions

This chapter needs a revision.

What are the minimal concepts which we should teach about multivariate statistics?

What are the basic methods which should at least be mentioned?

Summary

Summary

Time Table

Monday	Tuesday	Wednesday	Thursday	Friday
Basic Data Analysis (Ch. 1)	Regression (Ch. 2)	Regression	Regression: Discussion	Excerpts from Multivariate (Ch. 4)
Basic Data Analysis	Regression	Regression	Comparison (Ch. 3)	Excerpts from Multivariate
		Lunch Break		
Basic Data Analysis	Exercises	Unsupervised Exercises	Comparison	Exercises & Discussion
Exercises	Regression	Unsupervised Exercises	Exercises	Exercises & Discussion
		Afternoon Break		
Basic Data Analysis	Exercises	Unsupervised Exercises	Discussion	Discussion
Exercises & Discussion	Exercises & Discussion		Supplements from Ch. 03	

Issues to check: QQ-Plot, PP-Plot, Monte Carlo Bands	Issues to Check: Basic Diagnostics for Linear Regression	No Checks Today	Issues to Check: Stochastic Order; Shift Alternatives vs. General Differences	Issues to Check: Univariate vs. Multivariate Comparison

Summary

Open Issues

- Ch. 1 Distribution Diagnostic
  - Essentially stable.
- Ch. 2 Regression
  - Clarify Gauss-Markov theorem and role of dimension.
  - Can the chapter be cleaned up ?
- Ch. 3 Comparison (Still a placeholder)
  - What is an up-to-date discussion of the (non-shift) two sample case ?
  - Clean up.
- Ch. 4 Multivariate
  - Given the time limitations, is the current list of concepts sufficient ?
  - Discuss scaling issues, e.g. with respect to PCA.

Summary

Open Issues

- Ch. 1 Distribution Diagnostic
  - Essentially stable.
- Ch. 2 Regression
  - Clarify Gauss-Markov theorem and role of dimension.
  - Can the chapter be cleaned up ?
- Ch. 3 Comparison (Still a placeholder)
  - What is an up-to-date discussion of the (non-shift) two sample case ?
  - Clean up.
- Ch. 4 Multivariate
  - Given the time limitations, is the current list of concepts sufficient ?
  - Discuss scaling issues, e.g. with respect to PCA.

Summary

Open Issues

- Ch. 1 Distribution Diagnostic
  - Essentially stable.
- Ch. 2 Regression
  - Clarify Gauss-Markov theorem and role of dimension.
  - Can the chapter be cleaned up ?
- Ch. 3 Comparison (Still a placeholder)
  - What is an up-to-date discussion of the (non-shift) two sample case ?
  - Clean up.
- Ch. 4 Multivariate
  - Given the time limitations, is the current list of concepts sufficient ?
  - Discuss scaling issues, e.g. with respect to PCA.

Summary

Open Issues

- Ch. 1 Distribution Diagnostic
  - Essentially stable.
- Ch. 2 Regression
  - Clarify Gauss-Markov theorem and role of dimension.
  - Can the chapter be cleaned up ?
- Ch. 3 Comparison (Still a placeholder)
  - What is an up-to-date discussion of the (non-shift) two sample case ?
  - Clean up.
- Ch. 4 Multivariate
  - Given the time limitations, is the current list of concepts sufficient ?
  - Discuss scaling issues, e.g. with respect to PCA.

Summary

Open Issues

- Ch. 1 Distribution Diagnostic
  - Essentially stable.
- Ch. 2 Regression
  - Clarify Gauss-Markov theorem and role of dimension.
  - Can the chapter be cleaned up ?
- Ch. 3 Comparison (Still a placeholder)
  - What is an up-to-date discussion of the (non-shift) two sample case ?
  - Clean up.
- Ch. 4 Multivariate
  - Given the time limitations, is the current list of concepts sufficient ?
  - Discuss scaling issues, e.g. with respect to PCA.

Summary

Open Issues

- Ch. 1 Distribution Diagnostic
  - Essentially stable.
- Ch. 2 Regression
  - Clarify Gauss-Markov theorem and role of dimension.
  - Can the chapter be cleaned up ?
- Ch. 3 Comparison (Still a placeholder)
  - What is an up-to-date discussion of the (non-shift) two sample case ?
  - Clean up.
- Ch. 4 Multivariate
  - Given the time limitations, is the current list of concepts sufficient ?
  - Discuss scaling issues, e.g. with respect to PCA.

Summary

Open Issues

- Ch. 1 Distribution Diagnostic
  - Essentially stable.
- Ch. 2 Regression
  - Clarify Gauss-Markov theorem and role of dimension.
  - Can the chapter be cleaned up ?
- Ch. 3 Comparison (Still a placeholder)
  - What is an up-to-date discussion of the (non-shift) two sample case ?
  - Clean up.
- Ch. 4 Multivariate
  - Given the time limitations, is the current list of concepts sufficient ?
  - Discuss scaling issues, e.g. with respect to PCA.

Summary

Open Issues

- Ch. 1 Distribution Diagnostic
  - Essentially stable.
- Ch. 2 Regression
  - Clarify Gauss-Markov theorem and role of dimension.
  - Can the chapter be cleaned up ?
- Ch. 3 Comparison (Still a placeholder)
  - What is an up-to-date discussion of the (non-shift) two sample case ?
  - Clean up.
- Ch. 4 Multivariate
  - Given the time limitations, is the current list of concepts sufficient ?
  - Discuss scaling issues, e.g. with respect to PCA.

Summary

Open Issues

- Ch. 1 Distribution Diagnostic
  - Essentially stable.
- Ch. 2 Regression
  - Clarify Gauss-Markov theorem and role of dimension.
  - Can the chapter be cleaned up ?
- Ch. 3 Comparison (Still a placeholder)
  - What is an up-to-date discussion of the (non-shift) two sample case ?
  - Clean up.
- Ch. 4 Multivariate
  - Given the time limitations, is the current list of concepts sufficient ?
  - Discuss scaling issues, e.g. with respect to PCA.

Summary

Open Issues

- Ch. 1 Distribution Diagnostic
  - Essentially stable.
- Ch. 2 Regression
  - Clarify Gauss-Markov theorem and role of dimension.
  - Can the chapter be cleaned up ?
- Ch. 3 Comparison (Still a placeholder)
  - What is an up-to-date discussion of the (non-shift) two sample case ?
  - Clean up.
- Ch. 4 Multivariate
  - Given the time limitations, is the current list of concepts sufficient ?
  - Discuss scaling issues, e.g. with respect to PCA.

Summary

Open Issues

- Ch. 1 Distribution Diagnostic
  - Essentially stable.
- Ch. 2 Regression
  - Clarify Gauss-Markov theorem and role of dimension.
  - Can the chapter be cleaned up ?
- Ch. 3 Comparison (Still a placeholder)
  - What is an up-to-date discussion of the (non-shift) two sample case ?
  - Clean up.
- Ch. 4 Multivariate
  - Given the time limitations, is the current list of concepts sufficient ?
  - Discuss scaling issues, e.g. with respect to PCA.

Summary

References

See http://sintro.r-forge.r-project.org/.

Private Note

Document identification:

```
$Source: /u/math/j40/cvsroot/lectures/src/SIntro/Reisensburg_2009/CompStatBasic.tex,v $
$Revision: 1.16 $
$Date: 2009/07/25 20:09:50 $
$Name:  $

$Author: j40 $
```

LaTeX typeset: July 26, 2009