
Über die Möglichkeiten des Einsatzes von R in der Statistikausbildung

Bernhard Spangl,
Universität für Bodenkultur, Wien

Bad Doberan, 1. Okt. 2009

- * ... in der Lehre
 - * als Statistikprogramm
 - * zur Lösung von Übungsaufgaben
- * ... zur Generierung von
 - * Übungsaufgaben
 - * Tests
 - * Prüfungen

- * viele Studierende (~ 700 im WS, ~ 250 im SS)
- * früh im Studienplan (1., 2. oder 3. Semester), abhängig von der Studienrichtung
- * hauptsächlich Windows-Benutzer
- * nicht vertraut mit 'open source' oder 'mirrors' (oft sind es auch die Master-Studierenden nicht!)
- * keine weiteren Ressourcen (zeitlich, personell, EDV-Räume, etc.) vorhanden, um den Studierenden eine zusätzlichen Einführung in R zu geben
 \hookrightarrow wir stellen nur eine modifizierte Version von R zur Verfügung
 + Skriptum

- * Herunterladen & Installieren von R-2.9.2-win32.exe
- * benutzerdefinierte Installation, lediglich Auswahl von
 - * Main Files
 - * Compiled HTML Help Files
 - * Support Files for Package tcltk
 - * Message Translation
- * Installation zusätzl. Pakete, z.B. foreign, xlsReadWrite
- * Erzeugen des Verzeichnisses 'eigene R-Dateien'
- * Erzeugen eines entsprechenden Batch-Programms, um R zu starten
- * Erzeugen von `.First` & Default-Workspace `.RData`
- * Erzeugen von 'R-Ordner.zip'

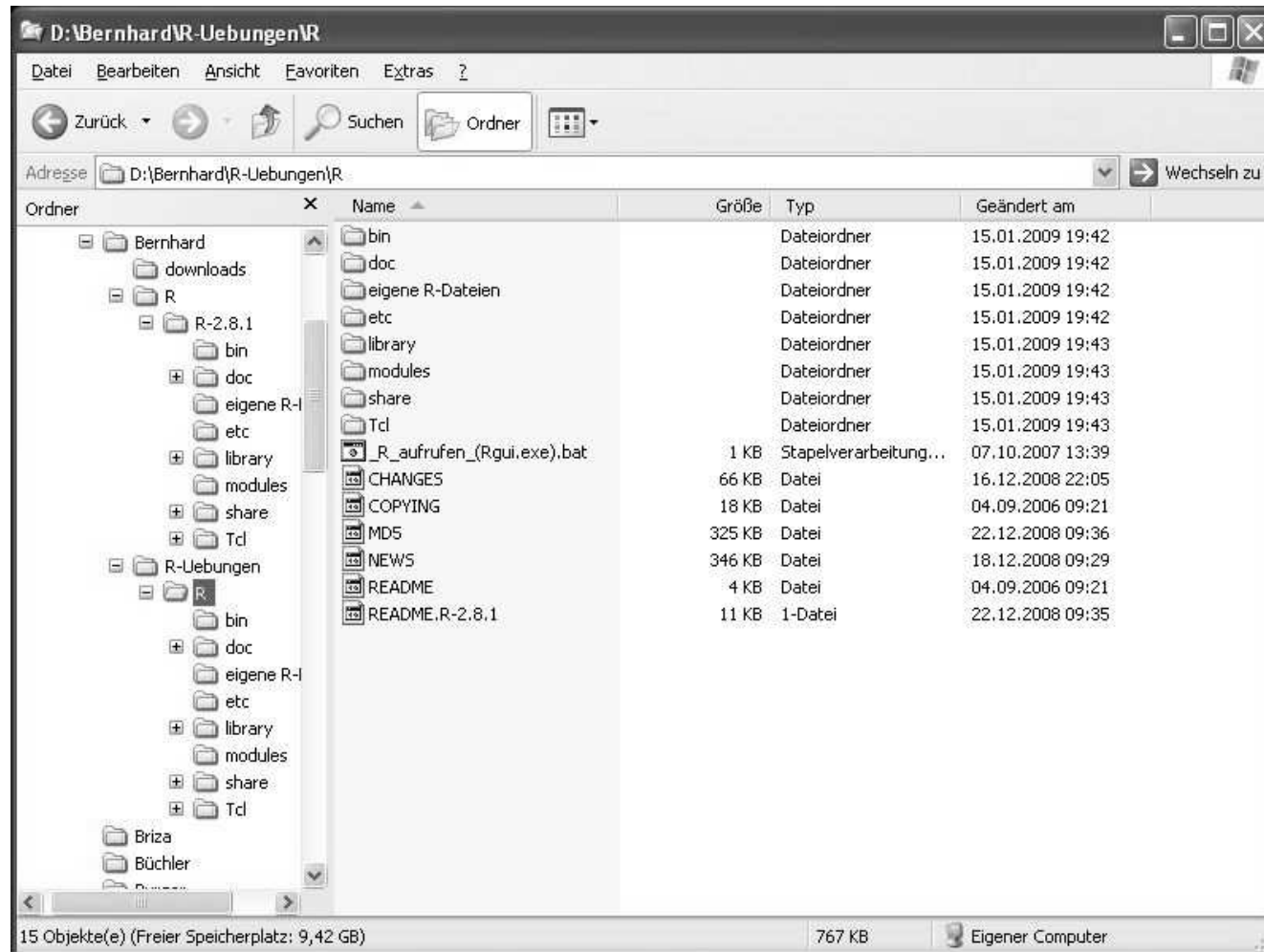
* Batch-Programm:

```
@echo off
echo Rgui wird gestartet ...
cd "eigene R-Dateien"
start ..\bin\Rgui.exe --sdi
```

* R-Funktion .First:

```
.First <- function () {
library(xlsReadWrite)
[...]
cat(getwd(), "\n")
[...]
}
```

Screenshot Explorer



ANOVA

Die entsprechenden Nullhypothesen für die Varianzanalyse lauten:

H_{Sorte} : Die Sorte hat keinen Einfluss auf den theoretischen mittleren Ertrag.

$H_{\text{Steinmehl}}$: Die Zugabe von Steinmehl zur Gülle hat keinen Einfluss auf den theoretischen mittleren Ertrag.

$H_{\text{Steinmehl:Sorte}}$: Es gibt keine Wechselwirkungen zwischen den beiden Faktoren Sorte und Steinmehl.

Die Alternativhypothesen lauten: "Sorte und Zugabe von Steinmehl zur Gülle haben Einfluss auf den theoretischen mittleren Ertrag und es gibt Wechselwirkungen zwischen den beiden Faktoren." Das Risiko 1. Art wird mit $\alpha = 0.05$ festgelegt.

```
> anova(lm(roggen$ertrag ~ roggen$steinmehl * roggen$sorte))
```

```
'+' ..... Modell ohne Wechselwirkungen
```

```
'*' ..... Modell mit Wechselwirkungen
```

```
'*' steht kurz für das Modell 'sorte + steinmehl + sorte: steinmehl'
```

Analysis of Variance Table

Response: roggen\$ertrag

	Df	Sum Sq	Mean Sq	F value	Pr(>F)	
roggen\$steinmehl	1	11.900	11.900	7.3731	0.014181	*
roggen\$sorte	2	73.091	36.545	22.6424	1.218e-05	***
roggen\$steinmehl:roggen\$sorte	2	25.556	12.778	7.9168	0.003414	**
Residuals	18	29.052	1.614			

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Die p-Werte für alle drei Hypothesen liegen unter dem gewählten Risiko 1. Art $\alpha = 0.05$, daher müssen alle Hypothesen verworfen werden. Es haben daher sowohl die Zugabe von Steinmehl zur Gülle als auch die Sorte Einfluss auf den Ertrag und es gibt Wechselwirkungen zwischen diesen beiden Faktoren.

Die Mittlere Quadratsumme des Fehlers (1.614) entspricht der Fehlervarianz s_e^2 .

* Vorteile:

- * Kombination der modifizierten R-Version + Skriptum gut von den Studierenden angenommen
- * portabel, einfach auf USB-Stick kopieren
- * aus didaktischer Sicht:
Die Studierenden hinterfragen ihre Resultate.
(Haben sie nie getan, als SPSS verwendet wurde!)

* Nachteil:

- * nur Anwenden von 'Kochrezepten', kein richtiges Verständnis/Erlernen von R

- * Konzept: Alexander Ploner
- * Verknüpfung von R und Latex
- * beliebig viele, individuelle Übungsangaben
- * seit etlichen Jahren im Einsatz
- * umfangreiche Funktionalität bereits vorhanden, musste allerdings erst geschaffen werden
- * mittlerweile überholt → Sweave

Screenshot Übungsaufgaben



The screenshot displays three overlapping windows from a Windows desktop environment:

- R Console:** Shows R code for loading data, creating an exam object, and writing it to a file. The code includes:

```
'help.start()' for an HTML browser interface to help.
Type 'q()' to quit R.

> setwd("D:/Be_university/Conferences/R-Herbsttagung2009/slides_spangl/")
> load("D:/Be_university/BOKU/Courses/EinfStat/stateinf_alex/AngStatR/.RData")
> studis <- read.table("uebBD09.txt",
+ sep=" ", colClasses=c("character","character"))
> setwd("D:/Be_university/BOKU/Courses/EinfStat/stateinf_alex/AngStatR/")
> source("data/StrKohlertrag2Fakt1.R")
> source("data/StrMaisertrag2Fakt1.R")
> source("data/StrRindergewicht2Fakt1.R")
> source("data/StrSchweinezunahme2Fakt1.R")
>
> testex <- createExam(list(StrKohlertrag2Fakt1, StrMaisertrag2Fakt1,
+ StrRindergewicht2Fakt1, StrSchweinezunahme2Fakt1),
+ label="Bad Doberan-2009", title="2. EDV-Beispiel", date="01.10.2009",
+ ID=studis[, 1], name=studis[, 2])
>
> writeExam(testex, "EDV2_2009_BadDoberan.tex")
```
- Adobe Reader:** Displays a PDF document titled "Bad Doberan 2009 2. EDV-Beispiel 01.10.2009". The document contains:
 - Header: "Bad Doberan 2009 2. EDV-Beispiel 01.10.2009"
 - Name: "MODER Karl", Mat.-nr.: "h1234567"
 - Instructions: "Bearbeiten Sie die folgenden Aufgaben mit R und geben Sie Ihre Resultate (versehen mit Name, Matrikelnummer, Buchstaben der Übungsgruppe, Beispielnummer und Beispiel-Code) bis zum jeweiligen Stichtag in den Übungen oder im Sekretariat des Instituts für Angewandte Statistik (Peter Jordan-Strasse 82, 1190) ab. Für die alternative elektronische Abgabemöglichkeit als Word-Datei im rtf-Format wird maximal 4 Seiten erlaubt. Verwenden Sie bitte die auf der Learning-Plattform zur Verfügung gestellte Vorlage! Die normale Abgabe gilt, die aus technischen Gründen notwendige Seitenanzahlrechnung nicht."
 - Attention: "1. Die Angaben enthalten einen Beispiel-Code (2 Buchstaben und eine Ziffer). Bitte legen Sie diesen Code unbedingt Ihrer Ausarbeitung bei! 2. Für die Ausarbeitung wird natürlich die Beantwortung der im Beispiel angeführten Fragen erwartet! 3. Der verwendete R-Output ist samt zugehöriger Befehlszeile (!) unverändert in die Ausarbeitung zu übernehmen!"
 - Example 1: "Stiere wurden ab einem Lebendgewicht von 340 kg gemästet, um den Einfluss der Einkreuzung von Fleischrassen auf die Mast- und Schlachtleistung festzustellen. Geprüft wurden zwei Gruppen, nämlich Fleischartiere (FV), und FV x Charolais (FV x CH). Zusätzlich wird auch zwischen Sommer- und Winterfütterung unterschieden. In der folgenden Tabelle ist das Mastendgewicht [in kg] angegeben (Code BG-1)."
 - Tables:

	FV				
Sommer	492	506	678	697	692
Winter	672	690	660	676	661

	FV x CH				
Sommer	554	540	658	616	627
Winter	728	697	616	634	661
 - Questions:
 - Stellen Sie das Mastendgewicht sowohl für die Rinderrasse als auch für die Fütterungsart durch gruppierte Kreisdiagramme dar. (1)
 - Testen Sie, ob die Faktoren Rinderrasse und Fütterungsart Einfluss auf das Mastendgewicht haben, bzw. ob es Wechselwirkungen zwischen Rinderrasse und Fütterungsart gibt. Formulieren Sie alle Hypothesen und die jeweiligen Entscheidungen als *parvo* Sätze. Das Risiko 1. Art wird mit $\alpha = 0.05$ festgelegt. Welches Modell (I, II oder gemischt) legen Sie zugrunde? (3)
 - Stellen Sie die Mittelwerte für alle Faktorstufenkombinationen in einem gemeinsamen Diagramm dar. Kann die Signifikanz bzw. Nicht-Signifikanz der Wechselwirkungen in Punkt b) aus diesem Diagramm abgelesen werden? (2)
 - Überprüfen Sie, ob die Voraussetzungen bezüglich der Varianzen erfüllt sind. (1)
 - Geben Sie einen Schätzwert für die Fehlervarianz σ^2 an. (1)
 - Überprüfen Sie schließlich mit Hilfe zweier einfacher Varianzanalysen einen möglichen Einfluss der Rinderrasse bzw. der Fütterungsart auf das Mastendgewicht. Was sagen Sie dazu? (2)
- GYM2:** Shows the LaTeX source code for the document, including:

```
\documentclass[10pt,german]{ASEExam}
\pagehead{Bad Doberan-2009}{2. EDU-Beispiel}{01.10.2009}
\begin{document}
\IDbox{MODER Karl}{h1234567}
{\Footnotesize
Bearbeiten Sie die folgenden Aufgaben mit R und geben Sie Ihre
Resultate (versehen mit Name, Matrikelnummer, Buchstaben der
Beispielnummer und {\bf Beispiel-Code}) bis zum
jeweiligen Stichtag in
%der \\'Ubung, Vorlesung
den \\'Ubungen oder im Sekretariat des
Instituts f\'ur Angewandte Statistik (Peter Jordan-Str\ass e 82, 1190) ab.
F\'ur die alternative elektronische Abgabem\oglichkeit als {\bf Word}-Datei
in {\bf rtf}-Format %(eine freiere Formatwahl w\urde uns
```

- * Verknüpfung von R und Latex mittels 'Sweave'
- * verschiedene, gleichwertige Testgruppen
- * Übungsgruppen an verschiedenen Wochentagen
- * aufbauend auf bereits vorhandener Funktionalität, plus eigene neue Funktionen
- * mit Lösungen!

Screenshot Tests



The screenshot shows a LaTeX editor window on the left and an Adobe Reader window on the right. The LaTeX editor displays the source code for a document titled 't060524.Rnw'. The code includes package declarations, page headers, and a task list. The task list contains three items: formulating a hypothesis, completing an ANOVA table, and making a decision based on the ANOVA results. The Adobe Reader window shows the rendered PDF document, which is a statistical test problem. The problem is titled '7. Test Statistik Mi., 31. Mai 06'. It includes a table of soil moisture data, two sub-questions (a and b), and a large ANOVA table. The solution for question (a) is provided in the LaTeX code, showing the hypothesis $H_0: \mu_1 = \mu_2 = \dots = \mu_k$ and the decision to reject H_0 at the 5% significance level.

```
\documentclass[10pt,german]{ASExam}
\usepackage{C:/Programme/R/rw2010/share/texmf/Sweave}
\pagehead{\Sexpr{test.nr}. Test}{Statistik}{\Sexpr{date}}
\begin{document}
\IDbox{}{\hspace{3cm}Gruppe:\hspace{1cm}}
\centerline{\bf AUSNAHMSWEISE OHNE UNTERLAGEN!}
\ex{1}

% Example: Bodenfeuchtigkeit
% Mead R. and Curnow R.N., Statistical Methods in Agriculture
% and experimental biology, Chapman and Hall, 1983, p. 46

In einem Feldversuch wird die Feuchtigkeit von \Sexpr{nlabelst1[1]}
verschiedenen Bodenarten [in Prozent] gemessen.
<<echo=FALSE, results=tex>>=
GroupedData1F(A1$df)
@

\begin{tasklist}
\item Formulieren Sie eine geeignete Hypothese (als ganzen Satz),
die sich mittels Varianzanalyse "uberpruefen"lassen. \acredit{1}

\item Ergaenzen Sie die Varianzanalysetabelle. \acredit{2}
\begin{center}
<<echo=FALSE, results=tex>>=
A1.aov <- aov(Feuchtigkeit~Boden, data=A1$df)
f.aov2latex.be(summary(A1.aov))
@
\end{center}
\item Entscheiden Sie "uber die Hypothese aus a) auf dem Niveau
\alpha=\Sexpr{alphaA1}$. Formulieren Sie Ihre Antwort wieder als
ganzen Satz, und geben Sie an, welches der \Sexpr{100*(1-alphaA1)}\%-Quantil
F_{f_1,f_2} aus der untenstehenden Tabelle Sie verwendet haben.
\acredit{1}
169, 0-1
```

7. Test Statistik Mi., 31. Mai 06

Name: _____ Mat.-nr.: _____ Gruppe: _____

AUSNAHMSWEISE OHNE UNTERLAGEN!

Beispiel 1

In einem Feldversuch wird die Feuchtigkeit von 4 verschiedenen Bodenarten [in Prozent] gemessen.

Boden	Feuchtigkeit					
E	18.2	14.9	10.3	13.4	9.2	12.7
B	10.2	17.0	9.6	13.1	15.3	16.7
C	2.6	13.1	14.8	2.6	6.1	13.6
A	11.2	4.8	13.4	16.2	7.7	9.8

a) Formulieren Sie eine geeignete Hypothese (als ganzen Satz), die sich mittels Varianzanalyse ueberpruefen laesst. (1)

b) Ergaenzen Sie die Varianzanalysetabelle. (2)

Faktor	SS	df	MS	F
Boden	92.95	?	30.98	?
Fehler	?	20	?	
Total	441.3	23		

c) Entscheiden Sie ueber die Hypothese aus a) auf dem Niveau $\alpha = 0.05$. Formulieren Sie Ihre Antwort wieder als ganzen Satz, und geben Sie an, welches der 95%-Quantile F_{f_1, f_2} aus der untenstehenden Tabelle Sie verwendet haben. (1)

	$f_2 = 17$	$f_2 = 18$	$f_2 = 19$	$f_2 = 20$	$f_2 = 21$	$f_2 = 22$	$f_2 = 23$	$f_2 = 24$
$f_1 = 2$	3.592	3.555	3.522	3.493	3.467	3.443	3.422	3.403
$f_1 = 3$	3.197	3.160	3.127	3.098	3.072	3.049	3.028	3.009
$f_1 = 4$	2.965	2.928	2.895	2.866	2.840	2.817	2.796	2.776
$f_1 = 5$	2.810	2.773	2.740	2.711	2.685	2.661	2.640	2.621

- * R ist kommandozeilen-orientiert (CLI)
- * graphische Benutzeroberfläche (GUI)
 - * R-Paket 'Rcmdr'
 - * jedoch nur eingeschränkte Funktionalität
- * Alternativen für Tests/Prüfungen:
 - * R-Paket 'exams'
 - * an WU Wien entwickelt und eingesetzt

Screenshot R-Commander



The screenshot displays the R-Commander interface with three main windows:

- Studentized Residuals (lm1):** A plot showing standardized residuals against t-quantiles. The y-axis ranges from -2 to 1, and the x-axis ranges from -2 to 0. A solid regression line is shown, along with dashed confidence intervals.
- roggen:** A data table with columns Duenger, Sorte, and Ertrag.
- lm(formula = Ertrag ~ Duenger * Sorte, data = roggen):** A summary of the linear model fit.

Duenger	Sorte	Ertrag	
1	mit	EhoKurz	42.0
2	mit	EhoKurz	45.2
3	mit	EhoKurz	46.3
4	mit	EhoKurz	44.7
5	ohne	EhoKurz	41.6
6	ohne	EhoKurz	43.7
7	ohne	EhoKurz	41.5
8	ohne	EhoKurz	40.8
9	mit	Motto	39.9
10	mit	Motto	42.4
11	mit	Motto	41.8
12	mit	Motto	40.4
13	ohne	Motto	42.6
14	ohne	Motto	43.7
15	ohne	Motto	41.8
16	ohne	Motto	42.4
17	mit	Kustro	39.5
18	mit	Kustro	41.3
19	mit	Kustro	39.6
20	mit	Kustro	41.9
21	ohne	Kustro	37.4
22	ohne	Kustro	39.0
23	ohne	Kustro	36.2
24	ohne	Kustro	37.4

```
lm(formula = Ertrag ~ Duenger * Sorte, data = roggen)

Residuals:
    Min       1Q   Median       3Q      Max
-2.5500 -0.8625 -0.1000  0.8125  1.8000

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)    44.5500    0.6352   70.133 < 2e-16 ***
Duenger[T.ohne] -2.6500    0.8983  -2.950 0.008569 **
Sorte[T.Kustro] -3.9750    0.8983  -4.425 0.000327 ***
Sorte[T.Motto]  -3.4250    0.8983  -3.813 0.001275 **
Duenger[T.ohne]:Sorte[T.Kustro] -0.4250    1.2704  -0.335 0.741847
Duenger[T.ohne]:Sorte[T.Motto]  4.1500    1.2704  3.267 0.004286 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.27 on 18 degrees of freedom
Multiple R-squared:  0.7919, Adjusted R-squared:  0.7341
F-statistic: 13.7 on 5 and 18 DF, p-value: 1.324e-05
```

Messages

```
[2] NOTE: The dataset roggen has 24 rows and 3 columns.
[3] NOTE: The dataset roggen has 24 rows and 3 columns.
```