

Grundlagen der Datenanalyse mit R
(R 1)

Sommersemester 2008

und

Statistik und Simulation mit R
(R 2)

Wintersemester 2008/2009

Dr. Gerrit Eichner
Mathematisches Institut der
Justus-Liebig-Universität Gießen
Arndtstr. 2, D-35392 Gießen, Tel.: 0641/99-32104

E-Mail: gerrit.eichner@math.uni-giessen.de

URL: www.math.uni-giessen.de/gerrit.eichner

Inhaltsverzeichnis

1	Einführung	1
1.1	Was ist R und woher kommt es?	1
1.2	Unsere Unix-Umgebung	1
1.3	Grundsätzliches zur Arbeit mit R unter Unix	4
1.3.1	Aufruf von R	4
1.3.2	Befehlseingabe und -ausführung in R	5
1.3.3	Benutzerdefinierte Objekte: Zulässige Namen, speichern und löschen	7
1.3.4	Die “command history”	8
1.3.5	R -Demonstrationen	8
1.3.6	Die R Online-Hilfe	8
1.3.7	Beenden von R	9
1.4	Grundsätzliches zur Arbeit mit R unter Windows	10
1.5	Einführungsliteratur	13
2	Datenobjekte: Strukturen, Attribute und elementare Operationen	14
2.1	Konzeptionelle Grundlagen	14
2.1.1	Atomare Strukturen/Vektoren	14
2.1.2	Rekursive Strukturen/Vektoren	14
2.1.3	Weitere Objekttypen und Attribute	15
2.1.4	Das Attribut „Klasse“ (“class”)	15
2.2	numeric -Vektoren: Erzeugung und elementare Operationen	15
2.2.1	Beispiele regelmäßiger Zahlenfolgen: seq() und rep()	16
2.2.2	Elementare Vektoroperationen	18
2.3	Arithmetik und Funktionen für numeric -Vektoren	19
2.3.1	Elementweise Vektoroperationen: Rechnen, runden, formatieren	19
2.3.2	Zusammenfassende und sequenzielle Vektoroperationen: Summen, Produkte, Extrema	20
2.3.3	“Summary statistics”: summary()	21
2.4	logical -Vektoren und logische Operatoren	23
2.4.1	Elementweise logische Operationen	23
2.4.2	Zusammenfassende logische Operationen	24
2.5	character -Vektoren und elementare Operationen	25
2.5.1	Zusammensetzen von Zeichenketten: paste()	25
2.5.2	Benennung von Vektorelementen: names()	26
2.5.3	Operationen für character -Vektoren: strsplit() , nchar() , substring() & abbreviate()	26
2.6	Indizierung und Modifikation von Vektorelementen: []	27
2.6.1	Indexvektoren	27
2.6.2	Zwei spezielle Indizierungsfunktionen: head() und tail()	28
2.6.3	Indizierte Zuweisungen	29
2.7	Faktoren und geordnete Faktoren: Definition und Verwendung	29
2.7.1	Erzeugung von Faktoren: factor()	30
2.7.2	Änderung der Levelsortierung und Zusammenfassung von Levels	31
2.7.3	Erzeugung von geordneten Faktoren: ordered()	31
2.7.4	Änderung der Levelordnung und Zusammenfassung von Levels bei geordneten Faktoren	31
2.7.5	Klassierung und Erzeugung von geordneten Faktoren: cut()	32
2.7.6	Tabellierung von Faktoren und Faktorkombinationen: table()	32
2.7.7	Aufteilung gemäß Faktor(en)gruppen sowie faktor(en)gruppierte Funktionsanwendungen: split() , apply() & ave()	32

2.8	Matrizen: Erzeugung, Indizierung, Modifikation und Operationen	34
2.8.1	Grundlegendes zu Arrays	34
2.8.2	Erzeugung von Matrizen: <code>matrix()</code>	34
2.8.3	Benennung von Spalten und Zeilen: <code>dimnames()</code> , <code>colnames()</code> , <code>rownames()</code>	35
2.8.4	Erweiterung mit Spalten und Zeilen: <code>cbind()</code> und <code>rbind()</code>	36
2.8.5	Matrixdimensionen und Indizierung von Elementen: <code>dim()</code> , <code>[]</code> , <code>head()</code> et al.	36
2.8.6	Einige spezielle Matrizen: <code>diag()</code> , <code>col()</code> , <code>lower.tri()</code> & Co.	38
2.8.7	Wichtige Operationen der Matrixalgebra	38
2.8.8	Zeilen- und spaltenweise Anwendung von Operationen: <code>apply()</code> & <code>sweep()</code>	39
2.8.9	Statistikspezifische Matrixfunktionen und Zeilen- bzw. Spaltenzusammen- fassungen: <code>cov()</code> , <code>colSums()</code> , <code>colMeans()</code> und Verwandte	40
2.8.10	Erzeugung spezieller Matrizen mit Hilfe von <code>outer()</code>	40
2.9	Listen: Konstruktion, Indizierung und Verwendung	41
2.9.1	Erzeugung und Indizierung: <code>list()</code> und <code>[[]]</code> , <code>head()</code> sowie <code>tail()</code> . .	41
2.9.2	Benennung von Listenelementen und ihre Indizierung: <code>names()</code> und <code>\$</code> . .	42
2.9.3	Komponentenweise Anwendung von Operationen: <code>lapply()</code> und <code>sapply()</code>	43
2.10	Data Frames: Eine Klasse „zwischen“ Matrizen und Listen	44
2.10.1	Indizierung von Data Frames: <code>[]</code> , <code>\$</code> , <code>head()</code> und <code>tail()</code> sowie <code>subset()</code>	44
2.10.2	Erzeugung von Data Frames: <code>data.frame()</code>	45
2.10.3	„Summary statistics“ und Struktur eines Data Frames: <code>summary()</code> und <code>str()</code>	46
2.10.4	Komponentenweise Anwendung von Operationen: <code>lapply()</code> und <code>sapply()</code>	47
2.10.5	Anwendung von Operationen auf nach Faktoren gruppierte Zeilen: <code>by()</code> .	47
2.10.6	„Öffnen“ und „Schließen“ von Data Frames und der Suchpfad: <code>attach()</code> , <code>detach()</code> und <code>search()</code>	48
2.11	Abfrage und Konversion der Klasse von Datenobjekten sowie Abfrage von <code>NA</code> , <code>NaN</code> , <code>Inf</code> und <code>NULL</code>	49
3	Import und Export von Daten	51
3.1	Datenimport aus einer Datei: <code>scan()</code> , <code>read.table()</code> und Co.	51
3.1.1	Die Funktion <code>scan()</code>	51
3.1.2	Die Beispieldaten „SMSA“	54
3.1.3	Die Funktion <code>read.table()</code> und ihre Verwandten	55
3.2	Datenexport in eine Datei	58
3.2.1	Die Funktionen <code>write()</code> , <code>sink()</code> und <code>write.table()</code>	58
4	Elementare explorative Grafiken	60
4.1	Grafikausgabe am Bildschirm und in Dateien	60
4.2	Explorative Grafiken für univariate Daten	60
4.2.1	Die Häufigkeitsverteilung diskreter Daten: Balken-, Flächen- und Kreis- diagramme sowie Dot Charts	61
4.2.2	Die Verteilung metrischer Daten: Histogramme, „stem-and-leaf“-Diagramme, Boxplots, „strip charts“ und Q-Q-Plots	64
4.2.3	Zur Theorie und Interpretation von Boxplots und Q-Q-Plots	67
4.3	Explorative Grafiken für multivariate Daten	69
4.3.1	Die Häufigkeitsverteilung bivariat diskreter Daten: Mosaikplots	69
4.3.2	Die Verteilung multivariat metrischer Daten: Streudiagramme	70
4.3.3	Die Verteilung trivariat metrischer Daten: Bedingte Streudiagramme („co- plots“)	73
4.3.4	Eine weitere Möglichkeit multivariater Darstellung: „Stars“	75

5	Wahrscheinlichkeitsverteilungen und Pseudo-Zufallszahlen	76
5.1	Die eingebauten Verteilungen	76
5.2	Bemerkungen zu Pseudo-Zufallszahlen in R	78
6	Definition neuer Funktionen	79
6.1	Ein Beispiel	79
6.2	Syntax der Funktionsdefinition	80
6.3	Spezifizierung von Funktionsargumenten	81
6.3.1	Argumente mit default-Werten	81
6.3.2	Variable Argumentezahl: Das „Dreipunktargument“	82
6.3.3	Zuordnung von Aktual- zu Formalparametern beim Funktionsaufruf	83
6.3.4	Rückgabewert einer Funktion	83
6.4	Kontrollstrukturen: Bedingte Anweisungen, Schleifen, Wiederholungen	84
7	Weiteres zur elementaren Grafik	86
7.1	Grafikausgabe	86
7.2	Elementare Plotfunktionen: <code>plot()</code> , <code>points()</code> , <code>lines()</code> & Co.	86
7.3	Die Layoutfunktion <code>par()</code> und Grafikparameter für <code>plot()</code> , <code>par()</code> und andere	87
7.4	Überschriften, Untertitel und Legenden	90
7.5	Einige (auch mathematisch) nützliche Plotfunktionen	92
7.5.1	Stetige Funktionen: <code>curve()</code>	92
7.5.2	Geschlossener Polygonzug: <code>polygon()</code>	92
7.5.3	Beliebige Treppenfunktionen: <code>plot()</code> in Verbindung mit <code>stepfun()</code>	92
7.5.4	Die empirische Verteilungsfunktion: <code>plot()</code> in Verbindung mit <code>ecdf()</code>	92
7.5.5	„Fehlerbalken“: <code>errbar()</code> im Package <code>Hmisc</code>	93
7.5.6	Mehrere Polygonzüge „auf einmal“: <code>matplot()</code>	94
7.6	Interaktion mit Plots	94
8	Para- und nicht-parametrische Inferenzstatistik in „klassischen“ Ein- und Zweistichprobenproblemen für metrische Daten	96
8.1	Auffrischung des Konzepts statistischer Tests	96
8.1.1	Motivation anhand eines Beispiels	96
8.1.2	Null- & Alternativhypothese, Fehler 1. & 2. Art	96
8.1.3	Konstruktion eines Hypothesentests im Normalverteilungsmodell	98
8.1.4	Der p -Wert	100
8.2	Konfidenzintervalle für die Parameter der Normalverteilung	102
8.2.1	Der Erwartungswert μ	102
8.2.2	Die Varianz σ^2	104
8.3	Eine Hilfsfunktion für die explorative Datenanalyse	104
8.4	Ein Einstichproben-Lokationsproblem	105
8.4.1	Der Einstichproben- t -Test	106
8.4.2	Wilcoxon's Vorzeichen-Rangsummentest	107
8.5	Zweistichproben-Lokations- und Skalenprobleme	110
8.5.1	Der Zweistichproben- F -Test auf Gleichheit der Varianzen	110
8.5.2	Der Zweistichproben- t -Test bei unbekanntem, aber gleichen Varianzen	111
8.5.3	Die Welch-Modifikation des Zweistichproben- t -Tests	112
8.5.4	Wilcoxon's Rangsummentest (Mann-Whitney U-Test)	113
8.6	Das Zweistichproben-Lokationsproblem für verbundene Stichproben	116
8.6.1	Die Zweistichproben- t -Tests bei verbundenen Stichproben	117
8.6.2	Wilcoxon's Vorzeichen-Rangsummentest für verbundene Stichproben	119
8.7	Tests auf Unkorreliertheit	120
8.7.1	Der Pearsonsche Test auf Unkorreliertheit	121
8.7.2	Der Spearmansche Rangtest auf Unkorreliertheit	122

8.8	Die einfache lineare Regression	125
8.9	Die Formelversionen der Funktionen für die Zweistichprobentests	127
8.10	Tests für Quotienten von Erwartungswerten der Normalverteilung	128
8.11	Testgüte und Bestimmung des Stichprobenumfangs für Lokationsprobleme	129
8.11.1	Der zweiseitige Einstichproben-Gaußtest	129
8.11.1.1	Herleitung der Gütefunktion	129
8.11.1.2	Interpretation und Veranschaulichung der Gütefunktion	130
8.11.1.3	Verwendungen für die Gütefunktion	132
8.11.1.4	Das Problem der unbekanntem Varianz	133
8.11.2	Der zweiseitige Einstichproben- <i>t</i> -Test	134
8.11.2.1	Herleitung der Gütefunktion	134
8.11.2.2	Verwendung der Gütefunktion	135
8.11.3	Der einseitige Einstichproben- <i>t</i> -Test	137
8.11.3.1	Gütefunktion: Herleitung, Eigenschaften und Veranschaulichung	137
8.11.3.2	Verwendung der Gütefunktion	138
8.11.4	Die Zweistichproben- <i>t</i> -Tests	139
8.11.4.1	Zwei verbundene Stichproben	140
8.11.4.2	Zwei unverbundene Stichproben	141
9	Zur Inferenzstatistik und Parameterschätzung für Nominaldaten	143
9.1	Bernoulli-Experimente mit <code>sample()</code>	143
9.2	Einstichprobenprobleme im Binomialmodell	144
9.2.1	Der exakte Test für die Auftrittswahrscheinlichkeit <i>p</i> : <code>binom.test()</code>	144
9.2.2	Der approximative Test für <i>p</i> : <code>prop.test()</code>	145
9.2.3	Konfidenzintervalle für <i>p</i>	146
9.3	Mehrstichprobentests im Binomialmodell	148
9.3.1	Zur Theorie der approximativen <i>k</i> -Stichproben-Binomialtests (Pearsons X^2 -Tests)	149
9.3.2	Zur Implementation der <i>k</i> -Stichproben-Binomialtests: <code>prop.test()</code>	151
9.3.2.1	Der Fall <i>k</i> = 2 Stichproben	151
9.3.2.2	Der Fall <i>k</i> ≥ 3 Stichproben	152
9.4	Testgüte und Bestimmung von Stichprobenumfängen im Binomialmodell	153
9.4.1	Einseitiger und zweiseitiger Einstichprobentest	153
9.4.2	Einseitiger und zweiseitiger Zweistichprobentest: (<code>power.prop.test()</code>)	154
9.5	Tests im Multinomialmodell	156
9.5.1	Multinomial-Experimente mit <code>sample()</code>	156
9.5.2	Der approximative χ^2 -Test im Multinomialmodell: <code>chisq.test()</code>	157
9.6	Kontingenztafeln	158
9.6.1	χ^2 -Test auf Unabhängigkeit zweier Faktoren und auf Homogenität	158
9.6.1.1	Zum Fall der Unabhängigkeit	159
9.6.1.2	Zum Fall der Homogenität	160
9.6.1.3	Der approximative χ^2 -Test auf Unabhängigkeit und der approximative χ^2 -Test auf Homogenität: <code>chisq.test()</code>	161
9.6.2	Fishers Exakter Test auf Unabhängigkeit zweier Faktoren	163
9.6.2.1	Die Implementation durch <code>fisher.test()</code>	163
9.6.2.2	Der Spezialfall der (2 × 2)-Tafel: Die Odds Ratio	164
9.6.3	Kontingenztafeln für <i>k</i> ≥ 2 Faktoren und ihre Darstellung: <code>xtabs()</code> & <code>fTable()</code>	166
9.6.3.1	Der Fall bereits registrierter absoluter Häufigkeiten	167
9.6.3.2	Der Fall explizit aufgeführter Levelkombinationen	169