

Skript
Statistik und Datenanalyse
WS 08/09

Prof. Dr. Gerald van den Boogaart

10. November 2008

Inhaltsverzeichnis

1	Statistische Daten und Modelle	1-1
1.1	Grundlagen	1-1
1.1.1	Begriffe	1-1
1.1.1.1	Was ist Statistik?	1-1
1.1.1.2	Was ist Datenanalyse?	1-1
1.1.1.3	Was ist Stochastik?	1-3
1.1.2	Die Rolle der Statistik in der Natur- und Sozialwissenschaften	1-3
1.1.3	Rollen der Stochastik in den Ingenieurwissenschaften	1-4
1.1.3.1	Themen, die im Rahmen der Vorlesung angesprochen werden	1-4
1.1.3.2	Komplexe Themen, die hier leider nicht behandelt werden können	1-5
1.2	Grundlegende Modelle der Statistik	1-6
1.2.1	Die Grundbegriffe	1-6
1.2.2	Grundgesamtheit	1-9
1.2.3	Stichprobe	1-10
1.2.4	Nichtrepräsentative statistische Daten	1-13
1.2.5	Statistisches Modell	1-13
1.2.5.1	Beobachtungen als Zufallsvariablen und Daten als ihre Realisierung	1-13
1.2.5.2	Verteilung	1-14
1.2.6	Verteilungsfunktion	1-15
1.2.6.1	Dichtefunktion	1-16
1.2.6.2	Kenngößen als Realisierung einer Zufallsvariablen	1-18
1.2.6.3	Wahrscheinlichkeitsmodell	1-19
1.2.6.4	Das statistische Modell	1-21
1.3	Datenmatrix	1-22
1.3.0.4.1	R: Zugriff auf Datenmatrizen in R	1-26
1.3.1	Abhängigkeit und Unabhängigkeit in der Datenmatrix	1-30
1.4	Skala	1-31
1.4.1	Der Begriff der Skala	1-31
1.4.2	Diskrete Skalen	1-31
1.4.2.1	Kategoriell	1-31
1.4.2.2	Ordinal als Sonderform der kategoriellen Skala	1-32
1.4.2.3	Nominalskala als fast kategorielle Skala	1-32
1.4.2.3.1	R: [1-32
1.4.2.4	Intervallskaliert	1-34
1.4.2.5	Anzahlen	1-34
1.4.3	Stetige Skalen	1-36
1.4.3.1	Reelle Skala	1-36
1.4.3.2	Relative Skala	1-37
1.4.4	Spezielle Skalen	1-37

1.5	Tafeln	1-38
1.5.1	Häufigkeitstafel	1-38
1.5.2	Kontingenztafel	1-39
1.5.3	Multivariate Kontingenztafeln	1-40
1.5.3.0.1	R: [.	1-42
2	Statistische Graphik und deskriptive Statistik	2-1
2.1	Vorbereitung	2-1
2.1.0.0.2	R: [.	2-1
2.1.1	Das Streudiagramm	2-5
2.1.2	Systematik der deskriptiven Methoden	2-7
2.2	Univariate Graphik und Beschreibung für stetige Daten	2-8
2.2.1	Punktdiagramm	2-10
2.2.1.1	Diskussion	2-10
2.2.2	Histogramm	2-10
2.2.2.0.1	Konfidenz: Genauigkeit des Histogramms	2-11
2.2.3	Normalverteilung als Referenzverteilung	2-16
2.2.3.1	Die Normalverteilung	2-16
2.2.3.2	Verteilungseigenschaften	2-16
2.2.4	Kenngrößen und Parameter	2-17
2.2.5	Lageparameter	2-19
2.2.5.1	Arithmetischer Mittelwert	2-19
2.2.5.1.1	R: Verwendung von lapply für die Berechnung mit Teildatensätzen	2-23
2.2.5.2	Theoretische Kenngrößen, Schätzung, Schätzfehler und Vertrauensbereiche	2-26
2.2.5.2.1	Schätzfehler: Mittelwert	2-27
2.2.5.3	Vertrauensbereiche/Konfidenzintervalle	2-28
2.2.5.3.1	Konfidenz: Konfidenzintervall des arithmetischen Mittelwertes	2-28
2.2.5.4	Geometrischer Mittelwert	2-29
2.2.5.4.1	R: Definition eigener Funktionen in R	2-31
2.2.5.4.2	Konfidenz: Konfidenzintervall des geometrischen Mittelwertes	2-31
2.2.5.5	Median	2-32
2.2.5.5.1	Konfidenz: Konfidenzintervall für den Median	2-33
2.2.5.6	Quantile	2-34
2.2.5.6.1	Konfidenz: Konfidenzintervall für die Quantile	2-36
2.2.5.7	Modalwert	2-38
2.2.6	Streuparameter	2-39
2.2.6.1	Interquartilsabstand	2-39
2.2.6.1.1	Konfidenz: IQR	2-40
2.2.6.2	Varianz und Standardabweichung	2-40
2.2.6.2.1	Konfidenz: Konfidenzintervall für die Varianz und Standardabweichung	2-47
2.2.6.3	Bereich	2-48
2.2.6.4	Relative Streuparameter	2-49
2.2.7	Boxplot oder Kastendiagramm	2-50
2.2.7.1	Aufbau eines Boxplot	2-50
2.2.7.1.1	R: Boxplots erzeugen	2-52
2.2.7.2	Interpretation des Boxplot	2-54
2.2.8	QQ-Plot	2-56
2.2.8.0.1	R: QQ-Plots	2-56

2.2.8.1	Interpretation von QQ-Plots	2-60
2.2.9	Konzept für relative Skala: Darstellung auf einer Log-Skala	2-64
2.3	Univariate Daten	2-67
2.3.0.0.1	R: Erzeugung einer Datentafel	2-67
2.3.1	Balkendiagramm/Barchart	2-69
2.3.2	Kenngrößen für die diskrete Daten	2-71
2.3.3	Kuchendiagramm/Piechart	2-71
2.4	Multivariate Graphik für stetig Daten	2-71
2.4.1	Streudiagramm	2-72
2.4.2	Kenngrößen für die stetige Abhängigkeit	2-72
2.4.3	log-Skala	2-76
2.4.4	Streudiagrammmatrix	2-76
2.4.5	Parallele Plots mit gleichem Koordinatensystem	2-76
2.5	Parallele Koordinaten	2-76
2.6	Graphik für gemischte Daten	2-83
2.6.1	Gesplittete Punktdiagramme	2-83
2.6.2	Gesplittete Boxplots	2-83
2.6.3	Gekerbte Boxplots	2-85
2.6.4	Labels und Farben	2-85
2.7	Multivariate Graphik für diskrete Daten	2-86
2.7.1	Gestapelte Balken	2-87
2.7.2	Parallele Balken	2-87
2.7.3	Mosaikplot	2-90
3	Statistische Tests	3-1
3.1	Der α -Niveau-Test	3-1
3.1.1	Einführendes Beispiel	3-1
3.1.2	Grundaufbau eines Tests	3-3
3.2	Weitere Grundbegriffe der Testtheorie	3-4
3.2.1	Interpretation als wissenschaftlicher Nachweis	3-4
3.2.2	Gütefunktion	3-4
3.2.3	Die Sümpfe des multiplen Testens	3-5
3.2.4	Der p-Wert eines Tests	3-7
3.2.5	Bezeichnungen	3-8
3.2.6	Testen auf einer Stichprobe: Der Ein-Stichproben-Gausstest	3-10
3.2.7	Alternative Alternativen	3-11
3.2.8	Was man zu einem Test wissen muss	3-14
3.2.9	Überblick über die Testsituationen	3-14
3.3	Die t-Tests und ihre Verwandten	3-16
3.3.1	Der Ein-Stichproben-t-Test	3-16
3.3.2	Relevante Verteilungen	3-17
3.3.3	Die Zwei-Stichproben-t-Tests	3-26
3.3.4	Welch-t-Test	3-28
3.3.5	Überprüfen der Varianzgleichheit	3-29
3.3.6	Gepaarter t-Test	3-30
3.3.7	Normalverteilungsmodelle mit mehreren Stichproben	3-30
3.4	Die Rangbasierten Tests	3-31
3.4.1	Wilcoxon-Rang-Summen-Test	3-31
3.4.2	Wilcoxon-Vorzeichen-Rang-Test	3-32
3.4.3	Weitere rangbasierte Tests	3-33
3.5	Anpassungstest	3-33
3.6	Multiples Testen	3-34
3.6.1	Diskussion	3-35
3.6.1.1	Verifikation vs. Falsifikation	3-35

3.6.1.2	Test vs. Entscheidung vs. Wissen	3-35
4	Regression	4-1
4.1	Allgemeines Regressionsmodell	4-1
4.1.1	Überblick über die Regressionsmodelle	4-1
4.2	Allgemeines lineares Modell	4-2
4.2.1	Definition	4-2
4.2.2	Beispiel lineare Regression	4-3
4.3	Statistik linearer Modelle	4-4
4.3.1	Ziele I	4-4
4.3.1.1	Beispiel: Transmissivität eines Grundwasserleiters	4-4
4.3.2	Design linearer Modelle	4-6
4.3.2.1	Aufsteigende Modellsequenzen	4-6
4.3.2.2	Problem: Auswahl des richtigen Modells	4-6
4.3.2.3	Lineare Regression	4-7
4.3.2.4	Multiple lineare Regression	4-9
4.3.2.5	Polynomiale Regression	4-9
4.3.2.6	Varianzanalyse/ANOVA	4-11
4.3.2.7	Multifaktorielle Varianzanalyse	4-13
4.3.2.8	Interaktion	4-13
4.3.2.9	Interaktion von Faktoren	4-14
4.3.2.10	Höhere Faktorinteraktionen	4-14
4.3.2.11	Geschachtelte Faktoren/nested Faktors	4-15
4.3.2.12	Lineare Modelle mit Regressoren und Faktoren	4-15
4.3.2.13	Faktor-Regressorinteraktion	4-16
4.3.2.14	Regressor-Regressorinteraktion	4-18
4.3.2.15	Ausblick: Zufallseffekte/random-effect-models	4-18
4.3.2.16	Aufsteigende Modellsequenzen	4-19
4.3.2.17	Anova-Tabellen I	4-19
4.3.2.18	Auswertung	4-19
4.3.2.19	Beispiel: Körpergröße	4-20
4.3.2.20	Beispiel: Körpergröße	4-21
4.3.3	Wiederholung: Modellvergleich	4-22
4.3.4	Erklärungskraft des Modells	4-23
4.3.4.1	Das Bestimmtheitsmaß R^2	4-23
4.3.4.2	Das wahre R^2	4-24
4.3.4.3	R^2 im Einsatz	4-24
4.3.4.4	Relatives R^2	4-25
4.3.4.5	Probleme mit R^2	4-25
4.3.4.6	Verbesserung durch R^2_{adj}	4-26
4.3.4.7	Vergleich: p -Wert und R^2	4-26
4.3.4.8	Konfidenzintervalle für R^2	4-26
4.3.5	Modellauswahl	4-27
4.3.5.1	Probleme des sequenziellen Testens in der Modell- auswahl	4-27
4.3.5.2	Optimalselektion	4-27
4.3.5.3	Vorwärtsselektion	4-27
4.3.5.4	Rückwärtsselektion	4-27
4.3.5.5	Kombinationsmethoden	4-28
4.3.5.6	Problem des multiplen Testens	4-28
4.3.6	Kontraste und Post-Hoc Methoden	4-28
4.3.6.1	Identifizierbarkeit von Parametern	4-28
4.3.6.2	Identifizierbarkeit von Kontrasten	4-28

4.3.6.3	Problem des multiplen Testens: Notwendigkeit von Post-Hoc-Tests	4-29
4.3.6.4	Das Bonferoni-Prinzip zur Korrektur von p-Werten und α -Niveaus beim multiplen Testen	4-29
4.3.6.5	Die Problemstellung der Post-Hoc-Tests	4-30
4.3.7	Post-Hoc-Tests	4-30
4.3.7.1	Einfacher Post-Hoc-Test: Tukeys HSD (Honest Significant Difference)	4-30
4.3.7.2	Äquivalenz zwischen Konfidenzintervallen und Tests	4-31
4.3.7.3	Problem der multiplen Konfidenzintervalle	4-31
4.3.7.4	Konfidenzintervalle nach Bonferoni	4-31
4.3.7.5	Konfidenzintervalle im Least Signifikant Difference Schema	4-32
4.3.7.6	Simultane Konfidenzintervalle nach Scheffé	4-33
4.4	Regressionsdiagnostik	4-33
4.4.1	Hebelwirkungen und Cook-Distanzen	4-33
4.4.1.1	Hebelwirkung/leverage	4-33
4.4.1.2	Cook-Distance/Einfluss	4-34
4.4.2	Robuste Regression	4-34
5	Logistische Modelle	5-1
5.1	Logistische Regression	5-1
5.2	Loglineare Modelle	5-1
A	Übungen	A-1
A.1	R Einführung	A-1
A.1.1	Grundlagen	A-1
A.1.2	Lineare Algebra	A-1
A.1.2.1	Rechnen mit Vektoren	A-1
A.1.2.2	Rechnen mit Matrizen	A-2
A.1.3	Datensätze	A-3
A.1.4	Programmierung	A-4
A.2	Univariate Graphik	A-5
A.2.1	Graphikbefehle	A-5
A.2.2	Ausblick auf anspruchsvolle Graphikgestaltung	A-6
A.2.3	Selbständige graphische Analyse von Datensätzen	A-6
A.3	Simulieren und Schätzen	A-7
A.4	Konfidenzschätzung und t-Verteilung	A-9
A.5	Tests I	A-9
A.6	Tests II	A-12
A.7	Tests und Graphiken selbständig anwenden	A-18
A.8	Lineare Modelle	A-18
B	Überblick über die Tests	B-1
B.1	Ein-Stichproben-Tests	B-1
B.1.1	Tests auf Verteilungsannahmen	B-1
B.1.2	Tests auf Lage	B-2
B.1.3	Tests auf Streuung	B-3
B.2	Zwei-Stichproben-Tests	B-3
B.2.1	Tests auf Verteilungsgleichheit	B-3
B.2.2	Tests auf Lagegleichheit	B-4
B.2.3	Tests auf Streuungsgleichheit	B-5
B.3	Gepaarte Tests	B-5
B.3.1	Tests auf Lage	B-5

B.3.2	Tests auf Abhängigkeit	B-6
B.4	Mehrstichproben Tests	B-7
B.4.1	Tests auf Gleichheit der Lage	B-7
B.4.2	Tests auf Gleichheit der Streuung	B-8
C	Daten	C-1
C.1	Acorn Size and Oak Tree Daten	C-1
C.2	Calcium Daten	C-1
C.3	Balance Daten	C-1
C.4	Kuckuckseier in den Nestern anderer Vögel	C-1
C.5	3D Sicht	C-1
C.6	Flohkäfer	C-1
C.7	Grundwasserleiter	C-2
C.8	Schlafdaten	C-2
C.9	Säugetiere	C-2
C.10	Weitre Daten von Statlib	C-2
C.10.1	Brustkrebsdaten	C-2
C.10.2	Chromatography	C-2
C.10.3	Kindergroessen	C-2
C.10.4	Rauchen und Krebs	C-2
C.11	Beispiele	C-2
D	Verzeichnisse	D-1