2 Einlesen von Datensätzen

2.1 Einlesen von SPSS-Datenfiles

SPSS bietet die beiden kompatiblen Datenformate *.sav* und *.por* zum Speichern und Einlesen von Dateien an. Daten dieses Formats können problemlos über den Menüpunkt `DATEI` → `ÖFFNEN` → `DATEN` unter Angabe des Dateinamens und der entsprechenden Dateinamenerweiterung eingelesen werden.

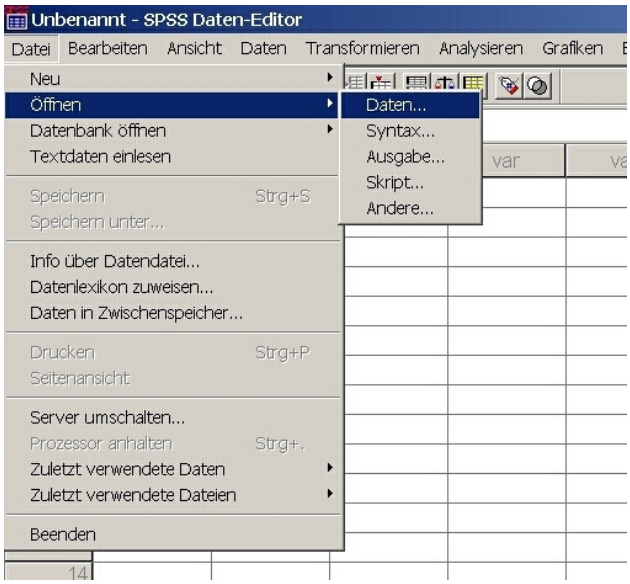


Abbildung 3: *Einlesen von Dateien*

Auf der CD-ROM finden Sie eine Reihe von Datensätzen im *.sav* und *.por*-Format, die Sie zu Testzwecken auf diese Weise einlesen können.

2.2 Einlesen von ASCII-Datenfiles

In der Praxis findet man Daten in vorformatierter SPSS-Form jedoch höchst selten. Häufiger liegen sie in SPSS-fremden Formaten vor, wobei vor allem dem ASCII-Format (American Standard Code for Information Interchange) eine große Bedeutung zukommt.

Das Einlesen solcher Formate gestaltet sich etwas schwieriger, da es keine einheitlichen Richtlinien hinsichtlich der Variablenseparation und der Trennzeichenformate gibt. SPSS verfügt deshalb über eine standardmäßige Abfrage der Syntax der Datenmatrix wie der Variablentypen. Die Auswahl des

ASCII-Formates *.dat* führt automatisch zur Öffnung des SPSS-Textimport-assistenten.

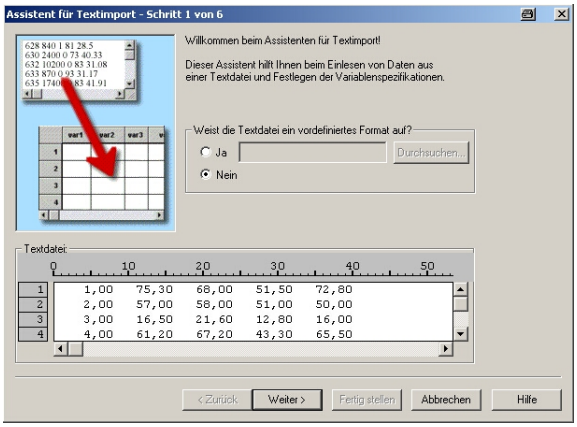


Abbildung 4: *Einlesen von ASCII-Daten*

In sechs Schritten kann der SPSS-Importfilter an die individuellen Merkmale der Daten angepasst werden[1]. Im Beispiel *kneipe.dat* sind die einzelnen Variablen durch Leerzeichen getrennt, wie in Abbildung 5 zu sehen ist.

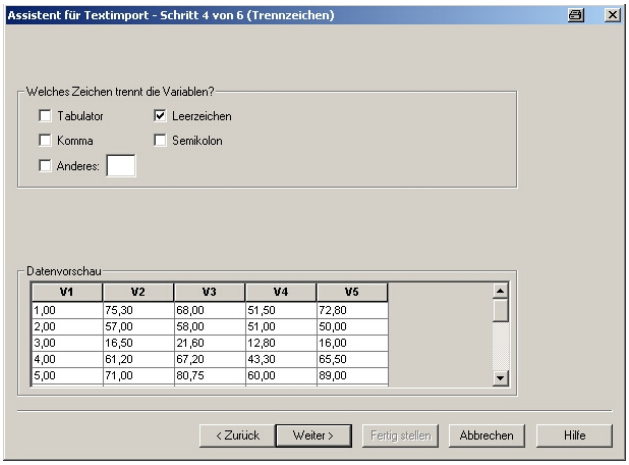


Abbildung 5: *Festlegen der Trennzeichnen*

SPSS bietet zusätzlich eine direkte Vorschau zur Verifizierung und Validierung der einzulesenden Daten und ihres Formats. Häufig auftretende Fehler in diesem Zusammenhang sind etwa Verwechslungen von Komma und Punkt bei der Dezimalstellentrennung oder multiples Auftreten von Zeichen, die zur Spaltenseparation vorgesehen sind, wie beispielsweise Leerzeichen, usw.

[1] Die einzelnen Schritte sind im Lernprogramm sehr übersichtlich dargestellt.

Kapitel 8

Klassifikation und Identifikation

8.1 Einführung in die Problemstellung und Übersicht

Klassifikation und Identifikation sind sogenannte *Q-Techniken* der multivariaten Statistik. Im Gegensatz zu den R-Techniken, die an den Objekten einer Grundgesamtheit beobachtbare Merkmale analysieren, dienen solche Q-Techniken der Strukturierung und Analyse der Objekte selbst. Q-Techniken

Die *Klassifikation* dient speziell der Bildung von Gruppen gleichartiger Objekte innerhalb einer Objektmenge. Das statistische Instrumentarium der Klassifikation ist die *Clusteranalyse*, die in Abschnitt 2 vorgestellt wird. Vermittels Clusteranalyse werden Objekte derart in *Klassen (Cluster)* eingeteilt, daß die Objekte eines jeden Clusters möglichst gleichartig (homogen) und die Cluster möglichst verschieden (heterogen) sind. Klassifikation Clusteranalyse

In Abschnitt 2.1 werden nun verschiedene *Klassifikationstypen* sowie Maße für die *Intraklassenhomogenität* und die *Interklassenheterogenität*, die auch zur Bewertung der *Güte einer Klassifikation* dienen, vorgestellt. Überblick

In den Abschnitten 2.2 und 2.3 werden dann *Konstruktionsverfahren* für die beiden wichtigsten Klassifikationstypen *(Partitionen und Hierarchien)* dargestellt.

Diskriminanzanalyse
Identifikation

Die *Diskriminanzanalyse*, vgl. Abschnitt 3, ist das Instrumentarium zur *Identifikation* von Objekten, d.h. zur Einordnung von „neuen" Objekten in vorgegebenen Objektklassen.

Überblick

Im Abschnitt 3.1 werden wir uns mit dem Fall beschäftigen, daß die Objektklassen disjunkt sind. Für solche Partitionen, vgl. Abschnitt 2, wird zunächst im Fall von p zugrunde liegenden normalverteilten Merkmalen der Zwei- und der Mehrgruppenfall der Diskriminanzanalyse behandelt. Außerdem werden in diesem Zusammenhang Verfahren zur Reduktion von Merkmalen behandelt; hier wird ausgehend von der Diskriminanzanalyse (unter Verwendung von Merkmalsinformation) überprüft, welche der dabei verwandten Merkmale nicht (signifikant) notwendig zur Identifikation von Objekten sind. Schließlich beschäftigen wir uns im Abschnitt 3.1 noch mit Verfahren der Diskriminanzanalyse bei beliebigen Grundgesamtheiten, die lediglich Distanzinformationen ausnutzen.

Im Abschnitt 3.2 wird dann noch ein Verfahren zur Einordnung neuer Objekte in die Klassen einer Hierarchie, vgl. Abschnitt 2.1, betrachtet, das ebenfalls lediglich Distanzinformationen benötigt.

8.2 Die Clusteranalyse

Strukturierung einer Objektmenge

Die Clusteranalyse dient der *Strukturierung* einer Menge von n Objekten $1, \ldots, n$ derart, daß diese Objekte in Klassen eingeteilt werden. Wir werden uns im folgenden hauptsächlich mit den wesentlichen Prinzipien und der konkreten Durchführung solcher Objektklassifikationen beschäftigen und nicht auf die entscheidungstheoretischen Grundlagen eingehen; man vergleiche hierzu etwa Bock (1973, Kap. III und IV), Degens (1978).

8.2.1 Allgemeine Prinzipien der Cluster-Analyse

Allgemeine Prinzipien der Cluster-Analyse

Voraussetzungen

Voraussetzung für die Durchführung einer Clusteranalyse für n Objekte $1, \ldots, n$ ist das Vorliegen einer quantitativen Datenmatrix

$$\mathbf{Y} = \begin{pmatrix} y_{11} & \cdots & y_{1p} \\ y_{21} & \cdots & y_{2p} \\ \vdots & & \vdots \\ y_{n1} & \cdots & y_{np} \end{pmatrix}$$

oder einer Distanzmatrix

$$\mathbf{D} = \begin{pmatrix} 0 & d(1,2) & d(1,3) & \dots & d_(1,n) \\ d(1,2) & 0 & d(2,3) & \dots & d_(2,n) \\ \vdots & & \ddots & \vdots & \\ d(1,n) & d(2,n) & d(3,n) & \dots & 0 \end{pmatrix}$$

für die n Objekte; man vergleiche auch die Ausführungen über Daten- und Distanzmatrizen im Abschnitt 1 des Kap. 7.

Ausgehend von solchen Informationen über eine Menge von n Objekten konstruiert jedes Klassifikationsverfahren (Clusteranalyseverfahren) eine *Klassifikation* Klassifikation

$$K = \{K_1, \dots, K_m\},$$

die aus den m *Klassen* $K_1, \dots, K_m$ von Objekten besteht. Jede dieser Klassen enthält mindestens eines und höchstens alle der n Objekte. Klassen einer Klassifikation

Das Klassifikationsverfahren selbst bildet aber erst den Abschluß einer *dreistufigen Vorgehensweise* bei der Clusteranalyse: Zunächst wird ein Klassifikationstyp gewählt und daran anschließend werden Bewertungskriterien für die Intraklassenhomogenität, die Interklassenheterogenität und die Klassifikationsgüte festgelegt. Erst dann wird ein Clusteranalyseverfahren bestimmt und die Klassifikation der n Objekte konkret durchgeführt. In diesem Abschnitt 2.1 werden nun zunächst die ersten beiden Stufen der Clusteranalyse behandelt. dreistufiges Vorgehen; Klassifikationstyp; Bewertungskriterien; Clusteranalyseverfahren

Die wichtigsten *Klassifikationstypen*, für die in Abschnitt 2.2 bzw. 2.3 auch Konstruktionsverfahren betrachtet werden, sind Partitionen und Hierarchien; sie sind Spezialfälle der Typen Überdeckung Quasihierarchie.

Man nennt eine Klassifikation $K = \{K_1, \dots, K_m\}$ eine *Überdeckung*, falls sich einzelne Klassen zwar überschneiden (gemeinsame Objekte besitzen) dürfen, jedoch keine Klasse vollständig in einer anderen enthalten ist: Überdeckung

$$\begin{aligned} K_i \cap K_j \quad &\notin \quad \{K_i, K_j\} \text{ für } i,j = 1,\dots,m \text{ mit } i \neq j \\ \Leftrightarrow K_i \cap K_j \quad &\neq \quad K_i \text{ und } K_i \cap K_j \neq K_j \end{aligned}$$

Beispiel

Beispiel:

In der *Abb. 8.1* werden 8 Objekte $1, 2, \ldots, 8$ graphisch durch Kreuze angedeutet. Die "Kreise" in der Abbildung geben 3 Objektklassen

$$K_1 = \{1, 2, 3, 4\},\ K_2 = \{3, 5, 6\},\ K_3 = \{2, 7, 8\}$$

an, die eine Überdeckung K für diese 8 Objekte bilden.

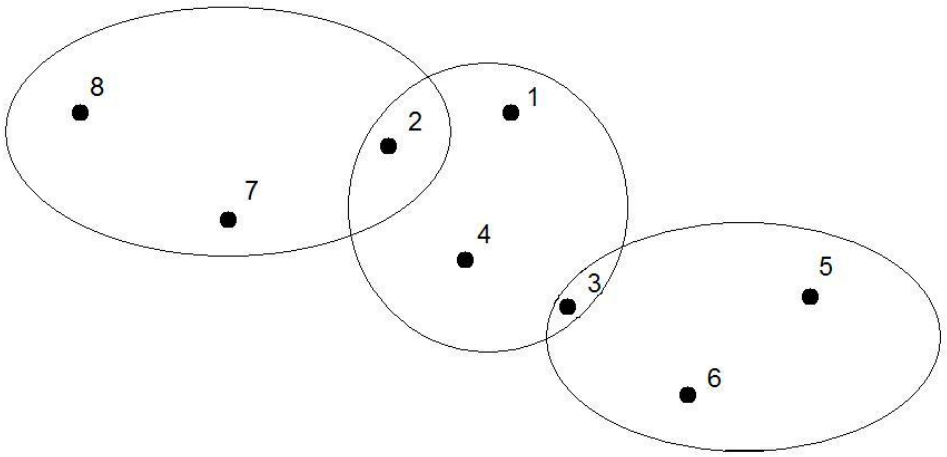


Abbildung 8.1: Graphische Veranschaulichung einer dreiklassigen Überdeckung für 8 Objekte

Quasihierarchie

Stammbaum der Stufen einer Quasihierarchie

Eine *Quasihierarchie* $K = \{K_1, \ldots, K_m\}$ ist eine Folge von Überdeckungen. Man kann sich eine Quasihierarchie in Form eines *"Stammbaumes"* vorstellen, dessen unterste Stufe die gröbste und dessen oberste Stufe die feinste in der Klassifikation K enthaltene Überdeckung darstellt. In den Klassen einer Stufe der Quasihierarchie sind dabei die Klassen einer darüberliegenden Stufe stets vollständig enthalten: Ist K_i eine Klasse aus der Quasihierarchie K, so gilt für die Vereinigung aller echten Teilklassen $K_j \overset{\subset}{\neq} K_i$ mit $K_j \in K \cup K_j \in \{\emptyset, K_i\}$.

Beispiel:

Die elf durch Punkte in *Abb. 8.2* veranschaulichten Objekte $1, 2, \ldots, 11$ sind, wie durch Kreise angedeutet, in 13 Klassen eingeteilt worden, die sich teilweise überschneiden oder vollständig in anderen Klassen enthalten sind. Beispiel

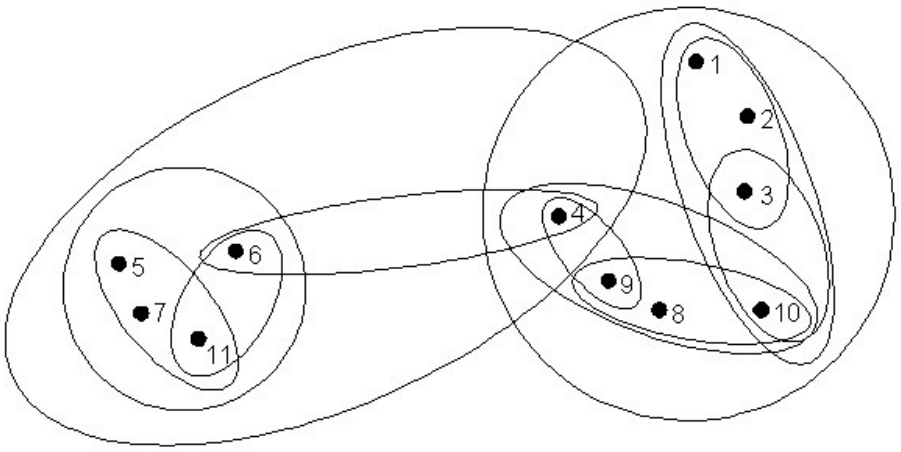


Abbildung 8.2: Graphische Veranschaulichung einer 13-klassigen Quasihierarchie für 11 Objekte

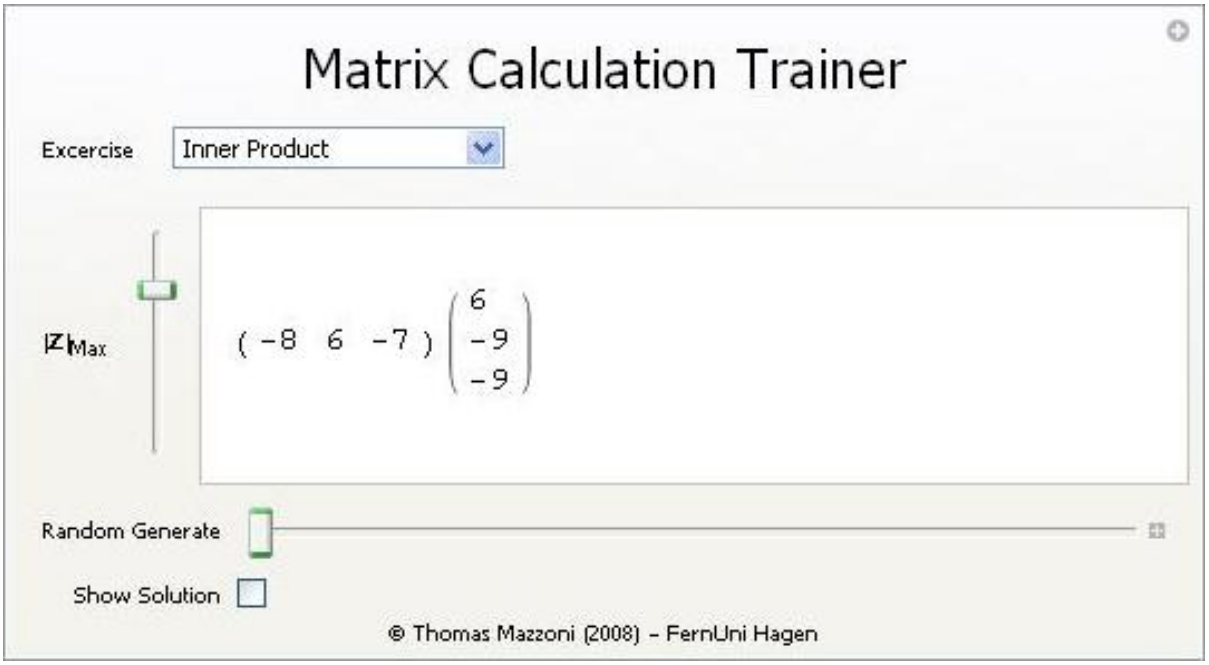
Aus diesen 13 Klassen

$$\begin{aligned}
K_1 &= \{4,9\},\ K_2 = \{3,10\},\ K_3 = \{6,11\}, K_4 = \{4,6\}.\\
K_5 &= \{1,2,3\},\ K_6 = \{8,9,10\},\ K_7 = \{5,7,11\},\\
K_8 &= \{4,8,9,10\},\ K_9 = \{5,6,7,11\},\ K_{10} = \{1,2,3,10\},\\
K_{11} &= \{4,5,6,7,11\},\ K_{12} = \{1,2,3,4,8,9,10\} \text{ und}\\
K_{13} &= \{1,2,3,4,5,6,7,8,9,10,11\}
\end{aligned}$$

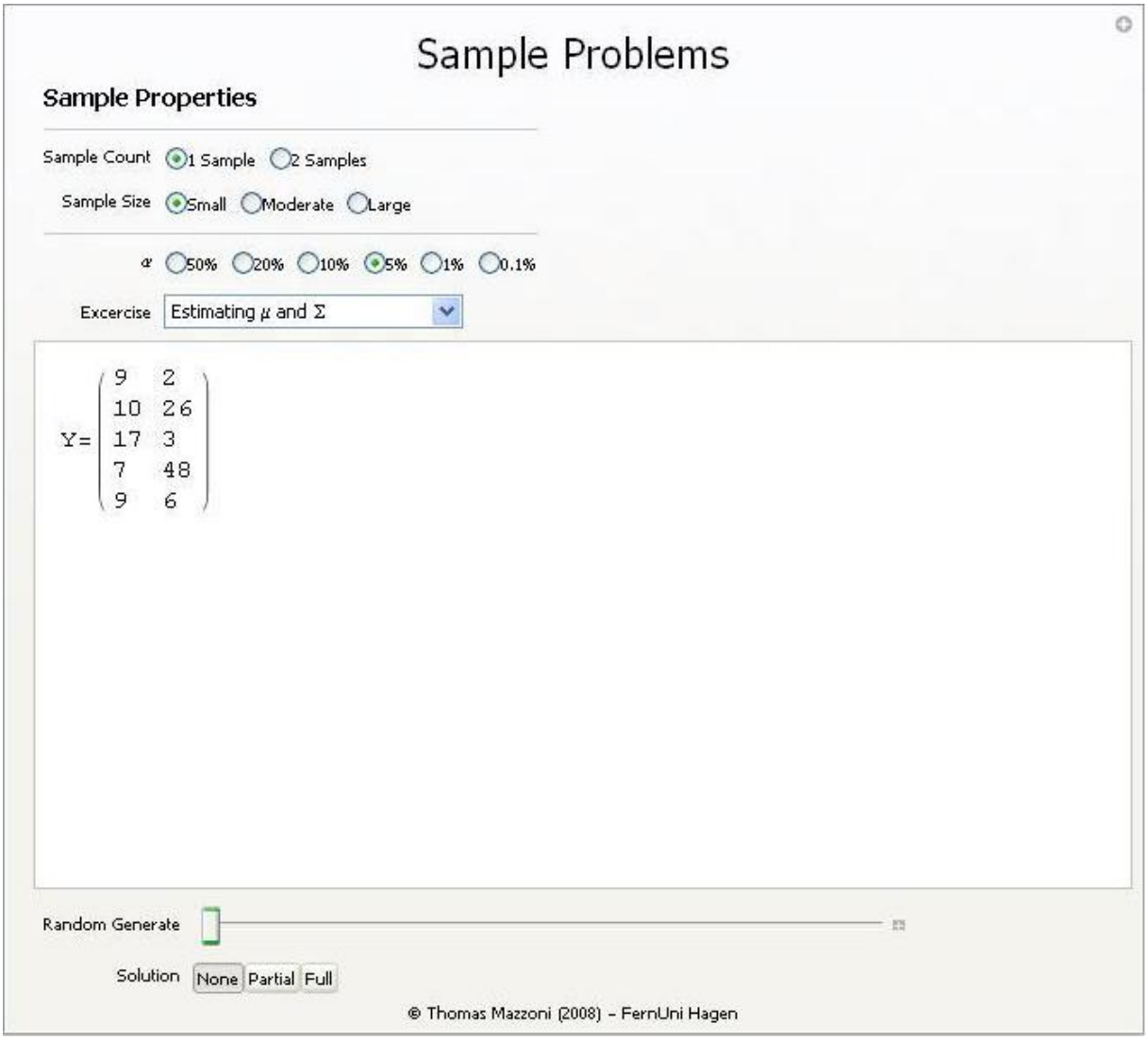
lassen sich verschiedene Quasihierarchien bilden. Zwei "Stammbäume" zu möglichen Quasihierarchien sind in den Abb. 3 und 4 angegeben. Die erste besteht aus 4 Stufen (Überdeckungen)

$$\begin{aligned}
K^0 &= \{K_{13}\},\ K^1 = \{K_{11}, K_{12}\},\ K^2 = \{K_4, K_8, K_9, K_{10}\},\ \text{und}\\
K^3 &= \{K_1, K_2, K_3, K_4, K_5, K_6, K_7\},
\end{aligned}$$

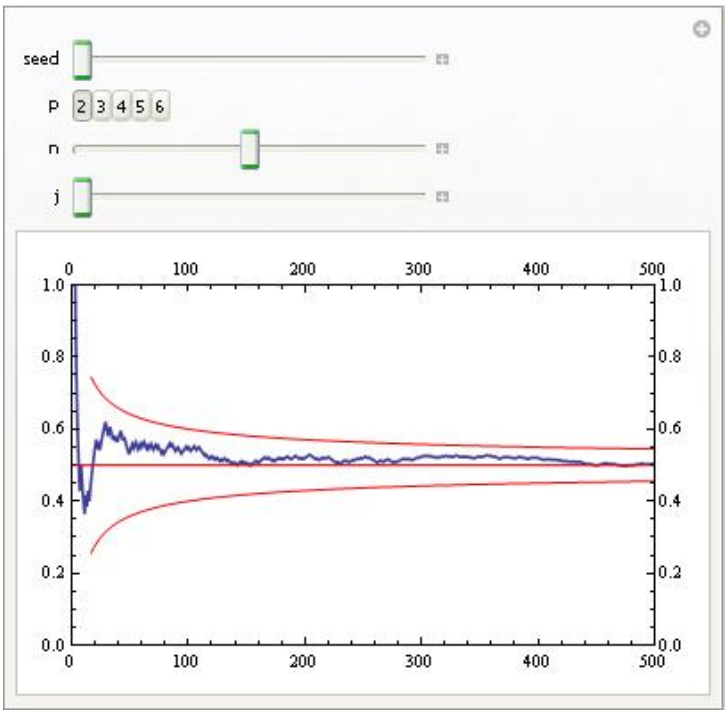
Lineare Algebra Trainer



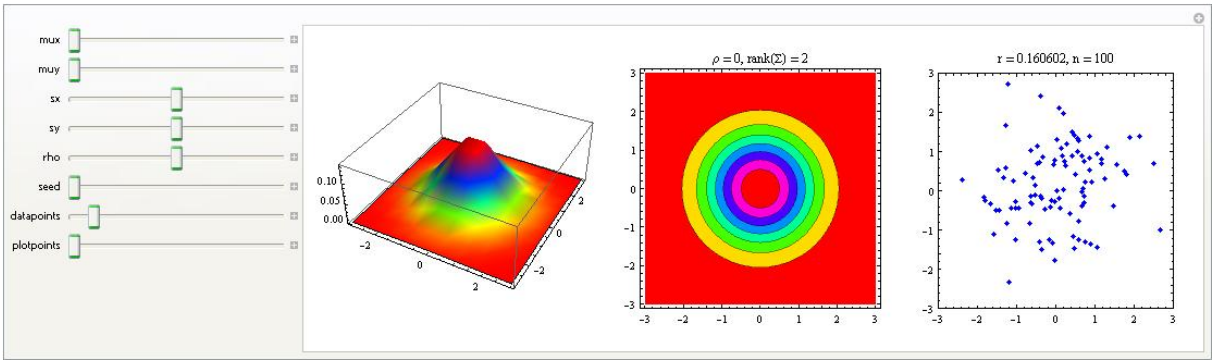
Stichprobentests



Relative Häufigkeiten



Bivariate Normalverteilung



Spurious Correlation