Einführung in die Statistik mit R

Bernd Weiler
syntegris information solutions GmbH
Neu Isenburg

Schlüsselworte
Statistik, R

Einleitung

Es ist seit längerer Zeit möglich statistische Berechnungen mit der freien Software R durchzuführen. Datenzugriffe sind über Schnittstellen zu fast allen Datenquellen (MS Office, Datenbanken, …) vorhanden. Seit einiger Zeit können mit R auch innerhalb der Oracle Datenbank Analysen erfolgen. Durch die R Integration können sich die Arbeitsabläufe verändern und Anwender haben die Möglichkeit Add-Hoc Analysen auf ihren Datenbeständen durchzuführen. Anwender ohne statistische Grundkenntnisse sind geneigt „ allgemein“ bekannte Kennzahlen und Methoden – z.B. Mittelwert, Korrelation, Regression – ohne Berücksichtigung der zugrunde liegenden Voraussetzungen zu verwenden.

Der Vortrag wendet sich an diese Anwender und ist eine Einführung in die Statistik mit Hilfe der Software R. Anhand von Beispielen soll gezeigt werden, welche grundlegenden statistischen Kennzahlen (Lageparameter) und Analysen (Methoden) unter Berücksichtigung des Skalenniveaus (nominal, ordinal, kardinal) der betrachteten Daten verwendet werden könnten/sollten. Anhand der Beispiele wird auf einige grundlegende Besonderheiten der Sprache R eingegangen. Es wird gezeigt, wie mit Hilfe einfacher graphischer Darstellung der betrachteten Daten Fehlspezifikationen (falsche Modellannahmen) erkannt bzw. das Ergebnis statistischer Analysen dargestellt werden.

Der Vortrag beschäftigt sich nicht eingehender mit der Sprache R innerhalb der Oracle Datenbank bzw. der Verbindung von der Desktopvariante von R mit Datenquellen.

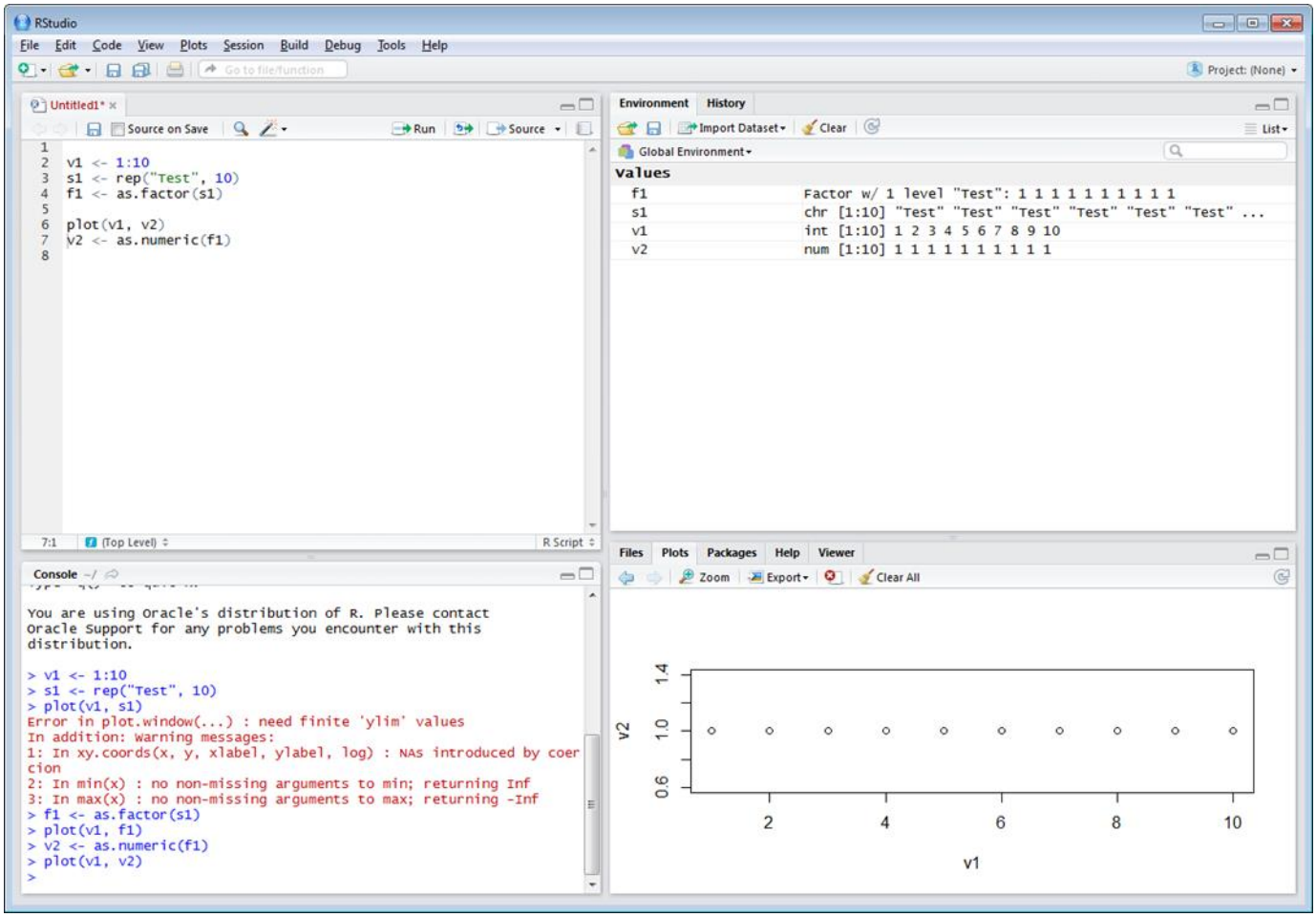
R

R ist eine freie Programmiersprache für statistisches Rechnen und statistische Grafiken. R ist Teil des GNU-Projekts und auf vielen Plattformen verfügbar. R gilt zunehmend als die Standardsprache für statistische Problemstellungen sowohl im kommerziellen als auch im wissenschaftlichen Bereich.

Benutzeroberflächen

R läuft in einer Kommandozeilenumgebung. Darüber hinaus gibt es mehrere grafische Benutzeroberflächen (GUI). Jede Oberfläche hat je nach Verwendungszweck ihre Vor- und Nachteile. Im Rahmen der Ausarbeitung dieses Vortrags wurden folgende Oberflächen verwendet:

- RGUI
- RStudio



Pakete

Der Funktionsumfang von R kann durch eine Vielzahl von Paketen erweitert und an spezifische statistische Problemstellungen aus diversen Anwendungsbereichen angepasst werden. Viele Pakete können dabei direkt aus einer über die R-Konsole abrufbaren Liste ausgewählt und automatisch installiert werden. Zentrales Archiv für diese Pakete ist das Comprehensive R Archive Network (CRAN).

Statistik

Statistik „ist die Lehre von Methoden zum Umgang mit quantitativen Informationen“ (Daten). Sie ist eine Möglichkeit, „eine systematische Verbindung zwischen Erfahrung (Empirie) und Theorie herzustellen“. Statistik wird einerseits als eigenständige mathematische Disziplin über das Sammeln, die Analyse, die Interpretation oder Präsentation von Daten betrachtet, andererseits als Teilgebiet der Mathematik, insbesondere der Stochastik, angesehen.
Untersuchungsgegenstand der Statistik sind Vorgänge, deren Resultate nicht mit Sicherheit vorhersehbar sind und die man daher als Zufallsexperimente bezeichnet. In diesem Sinne ist jede Messung, deren Resultate streuen, z.B. die Ausbildung der individuellen Körpergröße oder das Steueraufkommen einer Region ein Zufallsexperiment. Bemerkenswert ist nun aber, daß die Ergebnisse solcher Zufallsexperimente nicht regellos (chaotisch) anfallen. Sie lassen vielmehr Gesetzmäßigkeiten erkennen, die freilich nicht als einfache Wenn-Dann-Aussagen darstellbar sind:

Niemand weiß beispielsweise das Datum seines Todes. Eine Generation stirbt aber im Verlauf eines Jahrhunderts in ganz gesetzmäßiger Weise ab.
Die Menschen sind verschieden groß, ihre Körpergrößen sind aber nicht regellos verteilt. Wir wissen, daß Zwerge und Riesen nicht häufiger sind als Mittelwüchsige. Extreme Resultate des Wachstumsvorganges sind seltener als Durchschnittsresultate. Die Gesetzmäßigkeiten zufälliger Ereignisse geben dem Unvorhersehbaren einen Rahmen, machen Unsicherheit kalkulierbar. Durch geeignete Maßnahmen kann man Unsicherheit verringern. Das Fachgebiet der Statistik umfaßt einen Großteil der dazu verwendeten Methoden.

Teilgebiete der Statistik

Deskriptive Statistik

Die deskriptive Statistik (auch beschreibende Statistik oder empirische Statistik): Vorliegende Daten werden in geeigneter Weise beschrieben, aufbereitet und zusammengefasst. Mit ihren Methoden verdichtet man quantitative Daten zu Tabellen, graphischen Darstellungen und Kennzahlen. Bei einigen Institutionen ist wie bei der amtlichen Statistik oder beim sozio-oekonomischen Panel (SOEP) die Erstellung solcher Statistiken die Hauptaufgabe.

Die induktive Statistik

Die induktive Statistik (auch mathematische Statistik, schließende Statistik oder Inferenzstatistik): In der induktiven Statistik leitet man aus den Daten einer Stichprobe Eigenschaften einer Grundgesamtheit ab. Die Wahrscheinlichkeitstheorie liefert die Grundlagen für die erforderlichen Schätz- und Testverfahren.

Explorative Statistik

Die explorative Statistik (auch hypothesen-generierende Statistik, analytische Statistik oder Data-Mining): Dies ist methodisch eine Zwischenform der beiden vorgenannten Teilbereiche, bekommt als Anwendungsform jedoch zunehmend eine eigenständige Bedeutung. Mittels deskriptiver Verfahren und induktiver Testmethoden sucht sie systematisch mögliche Zusammenhänge (oder Unterschiede) zwischen Daten in vorhandenen Datenbeständen und will sie zugleich in ihrer Stärke und Ergebnissicherheit bewerten. Die so gefundenen Ergebnisse lassen sich als Hypothesen verstehen, die erst, nachdem darauf aufbauende, induktive Testverfahren mit entsprechenden (prospektiven) Versuchsplanungen sie bestätigten, als statistisch gesichert gelten können.

Betrachtungsgegenstand der Statistik

Der Betrachtungsgegenstand der Statistik sind (Zufalls-) Zahlen.

Skalenniveau

Quelle: http://de.wikipedia.org/wiki/Skalenniveau

Das Skalenniveau ist in der Empirie eine wichtige Eigenschaft von Merkmalen bzw. von Variablen. Je nach der Art eines Merkmals bzw. je nachdem, welche Vorschriften bei seiner Messung eingehalten werden können, lassen sich verschiedene Stufen der Skalierbarkeit unterscheiden:

Skalenniveau		log./math. Operationen	Beispiel
Nominalskala		=/≠	Geschlecht (Mann/Frau)
Ordinalskala		=/≠ ; </>	Schulnoten („sehr gut“ bis „ungenügend“)
Kardinalskala	Intervallskala	=/≠ ; </> ; +/−	Zeitskala (Datum)
	Verhältnisskala	=/≠ ; </> ; +/− ; ×/÷	Alter (0-99 Jahre)

Das Skalenniveau bestimmt

• die (mathematischen) Operationen, die mit einer entsprechend skalierten Variable zulässig sind. Dabei können Operationen, die bei Variablen eines bestimmten Skalenniveaus zulässig sind, grundsätzlich auch auf Variablen aller höheren Skalenniveaus durchgeführt werden. Ein auf einem bestimmten Niveau skalierbares Merkmal kann auf allen darunter liegenden Skalenniveaus dargestellt werden, jedoch nicht umgekehrt.

a) welche Transformationen mit entsprechend skalierten Variablen durchgeführt werden können, ohne Information zu verlieren bzw. zu verändern.

b) welche Information das entsprechende Merkmal liefert, welche Interpretationen Ausprägungen des entsprechenden Merkmals zulassen.

Das Skalenniveau gibt keine Auskunft darüber, ob eine Variable diskret (kategorial) oder stetig ist. Lediglich bei der Nominalskalierung kann man sicher sagen, dass das Merkmal nicht stetig sondern diskret ist.

Obwohl Skalenniveau und Anzahl der möglichen Ausprägungen unabhängige Konzeptionen darstellen, sind in der Praxis nominal- und ordinalskalierte Merkmale meist diskret und metrisch skalierte Merkmale meist stetig.

Methoden der statistischen Analyse

Die Schwerpunkte des Vortrages in Stichworten sind:

Lageparameter:

- Median
- Modus
- Mittelwert
- Quartile

- Standardabweichung (Varianz)
- Schiefe

Graphische Darstellungen

- Histogramm
- Box Plots
- Line Plots
- Scatter Plots

Methoden

- Korrelation
- Tabellen
- Varianzanalyse
- Regression

Kontaktadresse:
Bernd Weiler
syntegris information solutions GmbH
Hermannstraße 54-56, 63263 Neu-Isenburg
D-00000 Stadt

Telefon:	+49 (0) 6102 29 86 68
Fax:	+49 (0) 6102 55 88 062
E-Mail	Bernd.Weiler@syntegris.de
Internet:	http://www.syntegris.de/