§1 Keine Angst vor Daten

In der Statistik hat man mit Daten zu tun. Wir werden in diesem Abschnitt zunächst klären, welches Format statistische Daten üblicherweise besitzen und wie mit diesen Daten gearbeitet wird. Da statistische Daten oft sehr umfangreich sein können, ist es sinnvoll, derartige Daten in eigenen Files abzulegen. Wir werden in diesem Abschnitt zeigen, wie solche Files in einen eigenen Datenordner abgespeichert und von dort wieder aufgerufen werden können.

In vielen Fällen werden statistische Daten mit anderen Computerprogrammen erstellt bzw sollen anschließend mit anderen Computerprogrammen weiter verarbeitet werden. Wir werden in diesem Abschnitt zeigen, wie statistische Daten von Mathematica in MS-Excel exportiert bzw statistische Daten aus MS-Excel in Mathematica importiert werden.

1.1 Statistische Daten

Wir beginnen mit einer für die Statistik wichtigen Begriffsbildung, an die wir uns im folgenden streng halten:

1.1.1 Begriffsbildung: Statistische Daten werden stets in Form einer Matrix (Datenmatrix) angegeben. Eine Datenmatrix hat dabei stets die folgende Form:

In der **ersten Zeile** der Datenmatrix stehen die Namen der beobachteten Merkmale. Jede **weitere Zeile** der Datenmatrix enthält die an einem einzelnen Objekt - man spricht von einem Fall - ermittelten Werte dieser Merkmale.

Jede **Spalte** der Datenmatrix nennt man eine Variable. Die einzelnen Variablen entsprechen somit den beobachteten Merkmalen. Neben dem Namen des jeweiligen Merkmals enthält eine Variable jeweils die an den einzelnen Objekten ermittelten Werte dieses Merkmals - man nennt die Liste dieser Werte die zu diesem Merkmal gehörende Stichprobe.

Handelt es sich bei den Werten eines Merkmals um **Zahlen** (mit denen sich sinnvoll rechnen lässt), so spricht man von einem quantitativen Merkmal; handelt es sich bei diesen Werten hingegen um **Strings** (Zahlen, mit denen nicht sinnvoll gerechnet werden kann, fasst man dabei ebenfalls als Strings auf), so spricht man von einem qualitativen Merkmal. Fehlende Einträge werden in *Mathematica* mit "Null" gekennzeichnet.

Im folgenden Beispiel wird für einige ausgewählte Regionen des Staates Georgia das Pro-Kopf Einkommen, die Arbeitlosenrate, die Art der Region und die Anzahl der Einwohner in Form einer Datenmatrix angegeben:

Region	Einkommen	Arbeitslosenrate	Art	Einwohner
Bryan	10 826	4.8	Land	15 438
Bulloch	12 767	4.0	Land	43 125
Candler	14 214	6.2	Stadt	7744
Chatham	18 204	6.1	Stadt	216 935
Effingham	13 199	4.4	Land	25 687
Emanuel	12 682	6.9	Stadt	20 546
Evans	15 304	7.9	Stadt	8724
Jenkins	11 920	6.0	Land	8247
Liberty	11 470	Null	Land	52 745
Long	8922	3.2	Land	6202
Montgomery	13 357	7.0	Stadt	7163
Screven	12 610	6.3	Land	13 842
Tattnall	14 665	5.7	Stadt	17 722
Toombs	14 963	6.6	Stadt	24 072
Treutlen	11 684	Null	Land	5994
Wheeler	12 144	4.6	Land	4903

Abgesehen von der ersten Zeile mit den **Namen** der beobachteten Merkmale, entsprechen die Zeilen den ausgewählten Regionen. Bei den Merkmalen **Region** und **Art** handelt es sich um qualitative Merkmale; bei den Merkmalen **Einkommen**, **Arbeitslosenrate** und **Einwohner** handelt es sich um quantitative Merkmale. Für die Regionen Liberty und Treutlen ist die Arbeitslosenrate nicht bekannt. Bei der zum Merkmal **Arbeitslosenrate** gehörenden Stichprobe handelt es sich um die Liste {4.8, 4.0, 6.2, 6.1, 4.4, 6.9, 7.9, 6.0, , 3.2, 7.0, 6.3, 5.7, 6.6, , 4.6}.

Um mit dieser Datenmatrix tatsächlich arbeiten zu können, müssen wir sie in der üblichen Weise in *Mathematica* eingeben (wobei wir dieser Datenmatrix den Namen *georgia* geben):

```
georgia == {{Region, Einkommen, Arbeitslosenrate, Art, Einwohner}, {Bryan, 10 826, 4.8, Land, 15 438},
    {Bulloch, 12 767, 4.0, Land, 43 125}, {Candler, 14 214, 6.2, Stadt, 7744},
    {Chatham, 18 204, 6.1, Stadt, 216 935}, {Effingham, 13 199, 4.4, Land, 25 687},
    {Emanuel, 12 682, 6.9, Stadt, 20 546}, {Evans, 15 304, 7.9, Stadt, 8724}, {Jenkins, 11 920, 6., Land, 8247},
    {Liberty, 11 470, , Land, 52 745}, {Long, 8922, 3.2, Land, 6202}, {Montgomery, 13 357, 7., Stadt, 7163},
    {Screven, 12 610, 6.3, Land, 13 842}, {Tattnall, 14 665, 5.7, Stadt, 17 722},
    {Toombs, 14 963, 6.6, Stadt, 24 072}, {Treutlen, 11 684, , Land, 5994}, {Wheeler, 12 144, 4.6, Land, 4903}};
```

Mit dem Befehl TableForm lässt sich diese Datenmatrix in Form einer übersichtlichen Tabelle ausgeben (mit der Option TableSpacing lassen sich die Abstände zwischen den einzelnen Zeilen und Spalten einstellen):

```
TableForm[georgia, TableSpacing → {0.5, 3}]
```

```
Region       Einkommen   Arbeitslosenrate   Art      Einwohner
Bryan        10 826      4.8                Land     15 438
Bulloch      12 767      4.                 Land     43 125
Candler      14 214      6.2                Stadt    7744
Chatham      18 204      6.1                Stadt    216 935
Effingham    13 199      4.4                Land     25 687
Emanuel      12 682      6.9                Stadt    20 546
Evans        15 304      7.9                Stadt    8724
Jenkins      11 920      6.                 Land     8247
Liberty      11 470      Null               Land     52 745
Long         8922        3.2                Land     6202
Montgomery   13 357      7.                 Stadt    7163
Screven      12 610      6.3                Land     13 842
Tattnall     14 665      5.7                Stadt    17 722
Toombs       14 963      6.6                Stadt    24 072
Treutlen     11 684      Null               Land     5994
Wheeler      12 144      4.6                Land     4903
```

Mit dem Befehl Rest wird die erste Zeile dieser Datenmatrix (also jene Zeile, in der sich die Namen der einzelnen Variablen befinden) weggelassen:

```
TableForm[Rest[georgia], TableSpacing → {0.5, 3}]
```

Bryan	10 826	4.8	Land	15 438
Bulloch	12 767	4.	Land	43 125
Candler	14 214	6.2	Stadt	7744
Chatham	18 204	6.1	Stadt	216 935
Effingham	13 199	4.4	Land	25 687
Emanuel	12 682	6.9	Stadt	20 546
Evans	15 304	7.9	Stadt	8724
Jenkins	11 920	6.	Land	8247
Liberty	11 470	Null	Land	52 745
Long	8922	3.2	Land	6202
Montgomery	13 357	7.	Stadt	7163
Screven	12 610	6.3	Land	13 842
Tattnall	14 665	5.7	Stadt	17 722
Toombs	14 963	6.6	Stadt	24 072
Treutlen	11 684	Null	Land	5994
Wheeler	12 144	4.6	Land	4903

Mit dem Befehl Part lassen sich gezielt Fälle der Datenmatrix aufrufen. Der folgende Befehl ruft beispielsweise den 3-ten, 5-ten, 10-ten und 12-ten Fall der Datenmatrix georgia auf:

```
TableForm[Part[Rest[georgia], {3, 5, 10, 12}], TableSpacing → {0.5, 3}]
```

Candler	14 214	6.2	Stadt	7744
Effingham	13 199	4.4	Land	25 687
Long	8922	3.2	Land	6202
Screven	12 610	6.3	Land	13 842

Mit dem Befehl Part lassen sich aber auch gezielt Variable der Datenmatrix aufrufen. Der folgende Befehl ruft beispielsweise die 1-te, 2-te und 5-te Variable der Datenmatrix georgia auf:

```
TableForm[Part[georgia, All, {1, 2, 5}], TableSpacing → {0.5, 3}]
```

Region	Einkommen	Einwohner
Bryan	10 826	15 438
Bulloch	12 767	43 125
Candler	14 214	7744
Chatham	18 204	216 935
Effingham	13 199	25 687
Emanuel	12 682	20 546
Evans	15 304	8724
Jenkins	11 920	8247
Liberty	11 470	52 745
Long	8922	6202
Montgomery	13 357	7163
Screven	12 610	13 842
Tattnall	14 665	17 722
Toombs	14 963	24 072
Treutlen	11 684	5994
Wheeler	12 144	4903

In Verbindung mit dem Befehl Rest kann mit dem Befehl Part auch die zu einem Merkmal gehörende Stichprobe aufgerufen werden. So ruft etwa der folgende Befehl die zum 3-ten Merkmal (Arbeitslosenrate) gehörige Stichprobe der Datenmatrix georgia auf:

```
Part[Rest[georgia], All, 3]
```

```
{4.8, 4., 6.2, 6.1, 4.4, 6.9, 7.9, 6., Null, 3.2, 7., 6.3, 5.7, 6.6, Null, 4.6}
```

1.2 Schreiben und Lesen von Daten in Datenfiles

Mit dem Befehl Directory lässt sich ermitteln, welchen Ordner *Mathematica* gerade als aktuellen Ordner verwendet

(üblicherweise ist dies der Ordner "Eigene Dateien"). Möchte man einen anderen Ordner zum aktuellen Ordner machen, so kann dies mit dem Befehl SetDirectory erfolgen:

■ Directory[]

zeigt den gerade aktuellen Ordner an

■ SetDirectory[" *adresse* "]

macht den Ordner mit der Adresse *adresse* zum aktuellen Ordner (die Adresse eines Ordners findet man nach Öffnen dieses Ordners in der zugehörigen Adressleiste - es ist ratsam, diese Adresse mit Copy und Paste in den Befehl SetDirectory einzusetzen, wobei im dabei erscheinenden Dialogfeld "No" auszuwählen ist und auf die Anführungszeichen nicht vergessen werden darf).

Wir ermitteln den derzeit aktuellen Ordner:

```
Directory[]
```

```
C:\Dokumente und Einstellungen\Efrosinin\Eigene Dateien
```

Wir kommen überein, Datenfiles jeweils im Datenordner (dieser befindet sich im Ordner "Statistik") abzulegen. Dazu müssen wir diesen Datenordner zum aktuellen Ordner machen. Die Adresse dieses Ordners ist üblicherweise von Computer zu Computer verschieden und daher **individuell** einzugeben! Auf dem Computer, mit dem dieser Lehrgang erstellt wurde, geschieht dies mit dem Befehl

```
SetDirectory[$UserDocumentsDirectory]
```

```
C:\Dokumente und Einstellungen\Efrosinin\Eigene Dateien
```

Damit haben wir den Datenordner zum aktuellen Ordner erklärt:

```
Directory[]
```

```
D:\work\Lehre\Seminar\WTMS_SS2014
```

Um Schwierigkeiten zu vermeiden, führe man **jetzt** die beiden folgenden Schritte durch:

■ Man **ersetze** in der folgenden Zelle die Adresse durch die Adresse des Datenordners im gerade verwendeten Computer, **mache** diese Zelle zu einer initialisierenden Zelle (man aktiviere dazu diese Zelle und **wähle** anschließend im Menü Cell - Cell Properties die Option "Initialization Cell") und **evaluiere** diese Zelle:

```
SetDirectory[$UserDocumentsDirectory];
```

■ Man **ersetze** anschließend in allen mit pfad gekennzeichneten Notebooks des Ordners Statistik die unter dem jeweiligen Titel versteckte erste Zelle durch die eben geeignet abgeänderte Zelle. Wird daraufhin mit diesen Notebooks gearbeitet, so wird der Datenordner stets automatisch zum aktuellen Ordner erklärt.

Daten lassen sich mit dem Befehl Put in Dateien schreiben und mit dem Befehl Get aus Dateien einlesen:

■ Put[*daten*, *"datenfile"*] oder einfach *daten >> datenfile*

schreibt das Datenmaterial *daten* in das File *datenfile* des aktuellen Ordners. Wir kommen überein, jeweils nur ein einziges Datenmaterial *daten* in ein File zu schreiben und diesem File den Namen *datenfile* zu geben.

■ Get[*"datenfile"*] oder einfach << *datenfile*

liest das im File *datenfile* des aktuellen Ordners befindliche Datenmaterial *daten* ein.

Wir schreiben die Datenmatrix georgia unter Verwendung von Put (bzw >>) in das File georgiafile:

```
georgia >> georgiafile
```

Wir lesen die im File georgiafile gespeicherte Datenmatrix unter Verwendung von Get (bzw >>) ein:

```
<< georgiafile
```

```
{{Region, Einkommen, Arbeitslosenrate, Art, Einwohner},
 {Bryan, 10826, 4.8, Land, 15438}, {Bulloch, 12767, 4., Land, 43125},
 {Candler, 14214, 6.2, Stadt, 7744}, {Chatham, 18204, 6.1, Stadt, 216935},
 {Effingham, 13199, 4.4, Land, 25687}, {Emanuel, 12682, 6.9, Stadt, 20546},
 {Evans, 15304, 7.9, Stadt, 8724}, {Jenkins, 11920, 6., Land, 8247},
 {Liberty, 11470, Null, Land, 52745}, {Long, 8922, 3.2, Land, 6202},
 {Montgomery, 13357, 7., Stadt, 7163}, {Screven, 12610, 6.3, Land, 13842},
 {Tattnall, 14665, 5.7, Stadt, 17722}, {Toombs, 14963, 6.6, Stadt, 24072},
 {Treutlen, 11684, Null, Land, 5994}, {Wheeler, 12144, 4.6, Land, 4903}}
```

Wir lesen unter Verwendung der Befehle Get und Cases nur jene Fälle des Files georgiafile ein, welche bei der Variablen "Art" als Stadt gekennzeichnet sind und geben diese Daten mittels TableForm tabellarisch aus:

```
TableForm[Cases[<< georgiafile, {x_, y_, z_, Stadt, u_}], TableSpacing → {0.5, 3}]
```

```
Candler      14214   6.2   Stadt   7744
Chatham      18204   6.1   Stadt   216935
Emanuel      12682   6.9   Stadt   20546
Evans        15304   7.9   Stadt   8724
Montgomery   13357   7.    Stadt   7163
Tattnall     14665   5.7   Stadt   17722
Toombs       14963   6.6   Stadt   24072
```

1.3 Exportieren und Importieren von statistischen Daten

Ein in *Mathematica* vorliegendes statistisches Datenmaterial lässt sich mit dem Befehl Export in einer MS-Excel Datei ablegen. Ein in einer MS-Excel Datei abgelegtes statistisches Datenmaterial lässt sich mit dem Befehl Import in *Mathematica* einlesen (man beachte, dass die beiden Befehle Export und Import in dieser Form erst ab Version 6.0 unterstützt werden). Näheres dazu findet man im Tutorial tutorial/Importing and Exporting Data.

■ Export[*"daten.xls"*, *daten*]

legt das Datenmaterial *daten* im MS-Excel File *daten.xls* des aktuellen Ordners ab.

■ Import[*"daten.xls"*][[1]]

liest das im MS-Excel File *daten.xls* des aktuellen Ordners abgelegte Datenmaterial ein (da der Befehl Import das Datenmaterial mit einer zusätzlichen Klammer umgibt, ist der Ausdruck [[1]] erforderlich).

Wir legen das Datenmaterial georgia in der MS-Exel Datei *georgia.xls* ab:

```
Export["georgia.xls", georgia];
```

Wir lesen das in der MS-Excel Datei *stahl.xls* abgelegte Datenmaterial ein und nennen dieses Datenmaterial stahlxls. Da dieses Datenmaterial sehr umfangreich ist, listen wir von diesem Datenmaterial (unter Verwendung von

Take) nur die Namen der Variablen zusammen mit den ersten 4 Fällen auf:

```
stahlxls = Import["stahl.xls"][[1]];
TableForm[Take[stahlxls, 5], TableSpacing → {0.5, 3}]
```

```
Sorte   Kohlenstoff   Zugfestigkeit
A       42.           686.
A       42.           725.
B       46.           752.
B       46.           823.
```

Man beachte, dass dabei Strings *Mathematica*-intern stets mit Anführungszeichen " " versehen werden. Dies ist zu beachten, wenn aus Exel importierte Files weiter behandelt werden. Obwohl sich etwa das im File *stahlfile* abgelegte Datenmaterial stahl auf den ersten Blick nicht vom gerade importierten Datenmaterial stahlxls unterscheidet

```
TableForm[Take[<< stahlfile, 5], TableSpacing → {0.5, 3}]
```

```
Sorte   Kohlenstoff   Zugfestigkeit
A       42.           686.
A       42.           725.
B       46.           752.
B       46.           823.
```

sind diese beiden Datenmaterialien nicht! identisch

```
stahlxls === << stahlfile
```

```
False
```