

2a. Datenimport

Die wichtigste Frage vor Beginn einer Stata-Sitzung lautet:

Wie kommen meine Daten ins Programm ?

Je nach dem, in welcher Form die Daten vorliegen, gibt es hierfür mehrere Möglichkeiten.

1. Daten im Stata-Format

Ein vorhandenes Stata-Dokument, z.B. muscle.dta, das sich auf Laufwerk D: befindet (z.B. auf einem Datenstick) kann mit der Eingabe im Command-window:

use D:/Stata/muscle.dta

oder alternativ über *File >> Open* eingelesen werden.

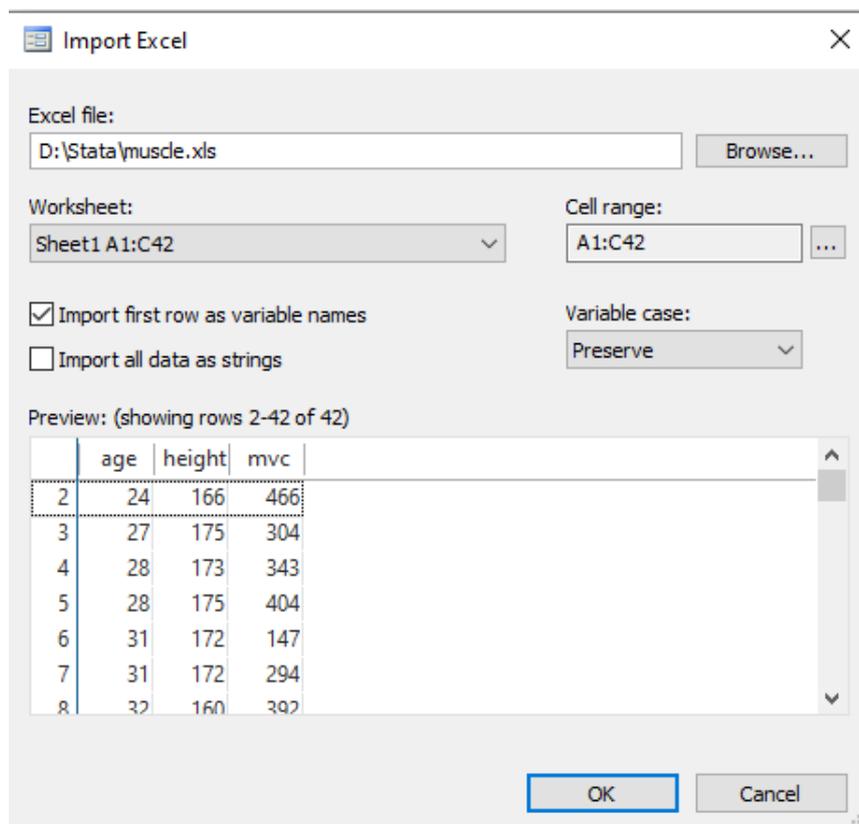
Wichtig ist die korrekte Pfad eingabe, damit Stata die Datei auch findet.

2. Daten in Excel-Tabelle

Die Daten befinden sich in einer Excel-Tabelle *.xls oder *.xlsx . Auch andere Kalkulationsprogramme (Planmaker, OpenOffice Calc ua.) erlauben normalerweise eine Speicherung im Excel-Format. Mit dem Kommando

import excel D:/Stata/muscle.xls , firstrow

werden die Daten in STATA eingelesen und die Variablennamen aus der ersten Zeile der Excel-Tabelle richtig positioniert (Option „firstrow“). Sind noch keine Variablennamen vorhanden, entfällt „firstrow“. Wichtig ist die korrekte Pfad eingabe. Das Dezimaltrennzeichen “.” oder “,” (3.5 oder 3,5) wird hier von STATA automatisch erkannt. Bei Menüsteuerung klickt man mit der Maus auf *File >> Import >> Excel* und erhält folgende Dialogbox:



Datenquelle für muscle.dta, fev.dta und lung1984.dta ist M. Bland: An introduction to medical statistics. Oxford Univ Press 2015

Mit „Browse“ sucht man nach der Excel-Datei, die man einlesen möchte und setzt bei „Import first row as variable names“ ein Häkchen. Mit „OK“ werden die Daten eingelesen und können mit *File >> Save as...* als Stata-Dokument gespeichert werden.

```
. sum var1 var2
```

Variable	Obs	Mean	Std. Dev.
var1	0		
var2	10	4.92	2.417437

3. Copy and Paste

Möchte man Daten aus einer Tabelle kopieren und in Stata einfügen ist darauf zu achten, vor dem Kopieren das Dezimaltrennzeichen auf "." (z.B. 3.5 statt 3,5) einzustellen, da

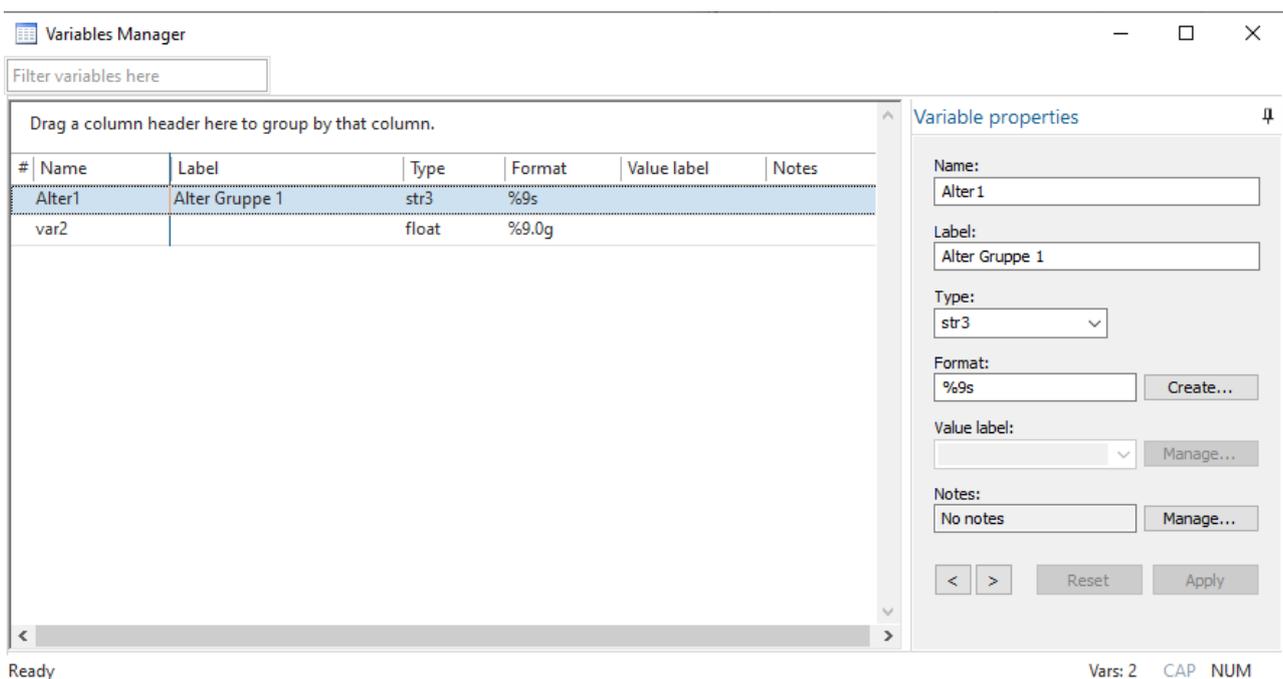
die Zahlen sonst als Text eingelesen werden. Hierzu gibt es in Tabellen häufig Möglichkeiten unter „Optionen“ oder „Einstellungen“. Für OpenOffice Calc z.B. geht man auf „Extras/Einstellungen/Spracheinstellungen/Sprachen“ und wählt dort unter „Gebietsschema: Englisch (GB)“. Man setzt noch ein Häkchen bei „Dezimaltrenntaste“ und mit OK erhält man statt der Kommas in der Tabelle Punkte. Jetzt markiert man den gesamten Datenbereich ohne Var-Namen in der Tabelle und kopiert mit Strg+C, wechselt dann in den Editor

	var1	var2
1	3,5	5.2
2	6,1	4.2
3	2,5	6.3
4	3,4	1.2
5	9,1	7.9
6	8,4	8.4
7	2,5	6.5
8	3,9	4.3
9	4,8	1.7
10	7,2	3.5

von Stata  und fügt die Daten mit Strg+V ein. Mit `File >> Save as...` speichert man die Daten als Stata-Datei. Die Namen der Var müssen noch vergeben werden.

4. Handeingabe bei wenig Daten

Durch Klick auf den Data Editor Button  oder Eingabe von `edit` in die Kommandozeile öffnet sich der Editor und man kann seine Daten von Hand eingeben, sofern der Umfang überschaubar ist. Beginnend in der ersten Spalte können jetzt die Daten eingegeben werden. Zahlenwerte erscheinen schwarz, Text (string) erscheint rot. Sind die Daten eingegeben, werden die Variablenamen neu vergeben, denn bis jetzt lauten sie noch `var1`, `var2`, usw. Hierzu öffnet man z.B. mit `Data >> Variables Manager` eine Dialogbox, in der verschiedene Änderungen vorgenommen werden können. `var1` wurde mit einem Komma, `var2` mit einem Punkt als Dezimaltrenner eingegeben. Stata erkennt `var1` als Text (String) und nur `var2` als Zahl. Mit Text können keine Rechenoperationen durchgeführt werden (Tabelle oben auf dieser Seite).



Zunächst ändert man die Variablennamen **var1** in **Alter1** und **var2** in **Alter2** mit Hilfe des Variables Manager oder alternativ auch mit den Kommandos:

rename var1 Alter1 und entsprechend **rename var2 Alter2** .

Ein Label kann ebenfalls eingefügt werden.

Mit **destring Alter1, replace dpcomma** wird das Komma bei der ersten Variablen in einen Punkt als Dezimaltrenner überführt und die Werte dann als Zahlen erkannt.

5. Daten aus einer Textdatei

Befinden sich die Daten in einer Tabelle, die eine Speicherung im Excel-Format *.xls oder *.xlsx nicht erlaubt oder deren Excel-Format in Stata nicht lesbar ist (z.B. ältere Formate) oder stehen die Daten in einer Textdatei zur Verfügung, so lassen sie sich meist in Textformat *.csv oder *.txt speichern. Zur Vermeidung von Problemen zwischen Komma als Dezimaltrennzeichen und Komma als Separation (csv) zwischen den Daten und zur besseren Übersicht sollte man wenn immer möglich den Tabulator als Trennzeichen zwischen den Zahlen verwenden. Das Kommando lautet:

import delimited D:/Stata/testdata.txt, delimiter(tab) varnames(1)

Mit **varnames(1)** werden die Variablennamen aus der ersten Zeile in den Spaltenkopf richtig positioniert.

Das Gleiche erreicht man im Menü mit *File >> Import >> Text data (delimited.....)*.

File to import:
D:\Stata\testdata.txt

Delimiter:
Tab

Use first row for variable names:
Always

Quote binding:
Loose

Floating point precision:
Use default

Text encoding:
Western (ISO Latin 1)

Variable case:
Lower

Quote stripping:
Automatic

Preview:

#	a	b
2	2,3	1,6
3	4,5	2,4
4	2,6	7,8
5	5,1	9,1

To change the data type for a column, right-click on the selected column and choose the appropriate type.

Buttons: OK, Cancel, Submit

Die Daten testdata.txt werden von Stata als String eingelesen (rot) und müssen noch mit **destring** in numerische Var überführt werden. Außerdem müssen noch die Var-Namen neu vergeben werden.

Einige Statistiklehrbücher stellen Daten kostenfrei zum Download zur Verfügung, z.B.:

Martin Bland: An Introduction to Medical Statistics. Oxford University Press 2015.
www-users.york.ac.uk/~mb55/intro/introcon.htm

Zur Verwendung siehe auch:
<https://www-users.york.ac.uk/~mb55/datasets/datasets.htm#intro>

Ein besonderes Textformat ist das dct-Format. Die Datei muscle.dct beispielsweise ist in diesem Format gegeben und kann eingelesen werden mit **infile using D:\Stata\muscle.dct**

oder im Menü *File >> Import >> Text data in fixed format with a dictionary*.

Das „dictionary“ ist ein Vorspann, der die Variablen und deren Label beinhaltet. Man beachte die Lage der Klammern { und }. Für muscle.dct z.B.:

```
dictionary {  
age "Age (years)"  
height "Height (cm)"  
mvc "Max voluntary contraction, quadriceps muscle (newtons)"  
}  
  24    166    466  
  27    175    304  
  28    173    343
```

Text-Daten (Strings) oder Zahlen mit Komma als Dezimaltrennzeichen sind nicht zugelassen zum Import von dct-Dateien. Zahlen mit "." als Dezimaltrennzeichen sind zugelassen.

Anmerkung: Beide Schrägstriche "\ " oder "/" bei der Pfadangebe sind möglich, also: **D:\Stata\muscle.dct** oder **D:/Stata/muscle.dct**.

STATA - Kommandos für Datenimport

use

edit

rename

import excel

import delimited

infile using

destring v1, replace dpcomma