GESStabs

Gewichtung

Gesellschaft für Software
in der Sozialforschung mbH

Waterloohain 6 - 8
22769 Hamburg
Tel.: 040 - 853 753 - 0
Fax: 040 - 853 753 - 33
www.gessgroup.de

GESS HowTo: Gewichten

Die Basis von Tabellierungen sind häufig Stichproben aus einer Grundgesamtheit, von der zentrale Parameter bekannt sind, wie z.B. die Geschlechts- oder Altersverteilung oder z.B. die geographische Verteilung auf Bundesländer. Die Stichproben sind ab und zu alles Andere als perfekt, was die Abbildung dieser Verteilungen betrifft. Dann entsteht der Wunsch nach einer Gewichtung.

GESS tabs kann passende Gewichte ermitteln, die einen Datensatz an eine oder mehrere gegebene Randverteilungen anpasst. Hierfür muss man lediglich im Script mitteilen, wie die einzelnen Ausprägungen einer besetzt sein sollen. Die Angaben erfolgen in Prozent.

Wir verwenden den Datensatz, den wir schon im Tutorial benutzt haben. Statt EXAMPLE.TAB verwenden wir hier eine leicht abgewandelte Steuerdatei **WEIGHT.TAB**.

Für ein erstes Beispiel soll die Randverteilung nach dem Geschlecht angepasst werden, und die Vorgabe sei, die Männer auf 48%, die Frauen auf 52 % zu gewichten.

Das Geschlecht ist in der Variablen „geschl“ abgelegt. Wie der entsprechende Abschnitt aus example.inc zeigt

```
variable geschl = * 1 labels
1 "Männlich"
2 "Weiblich"
;
```

ist Männlich mit 1, Weiblich mit 2 kodiert. Eine Gewichtungsvorgabe wird mit dem WEIGHTCELLS Statement formuliert. Wir legen eine neue Include-Datei mit Namen WEIGHT.INC an und tragen dort ein:

```
weightcells geschl =
1 : 48%
2 : 52%
;
```

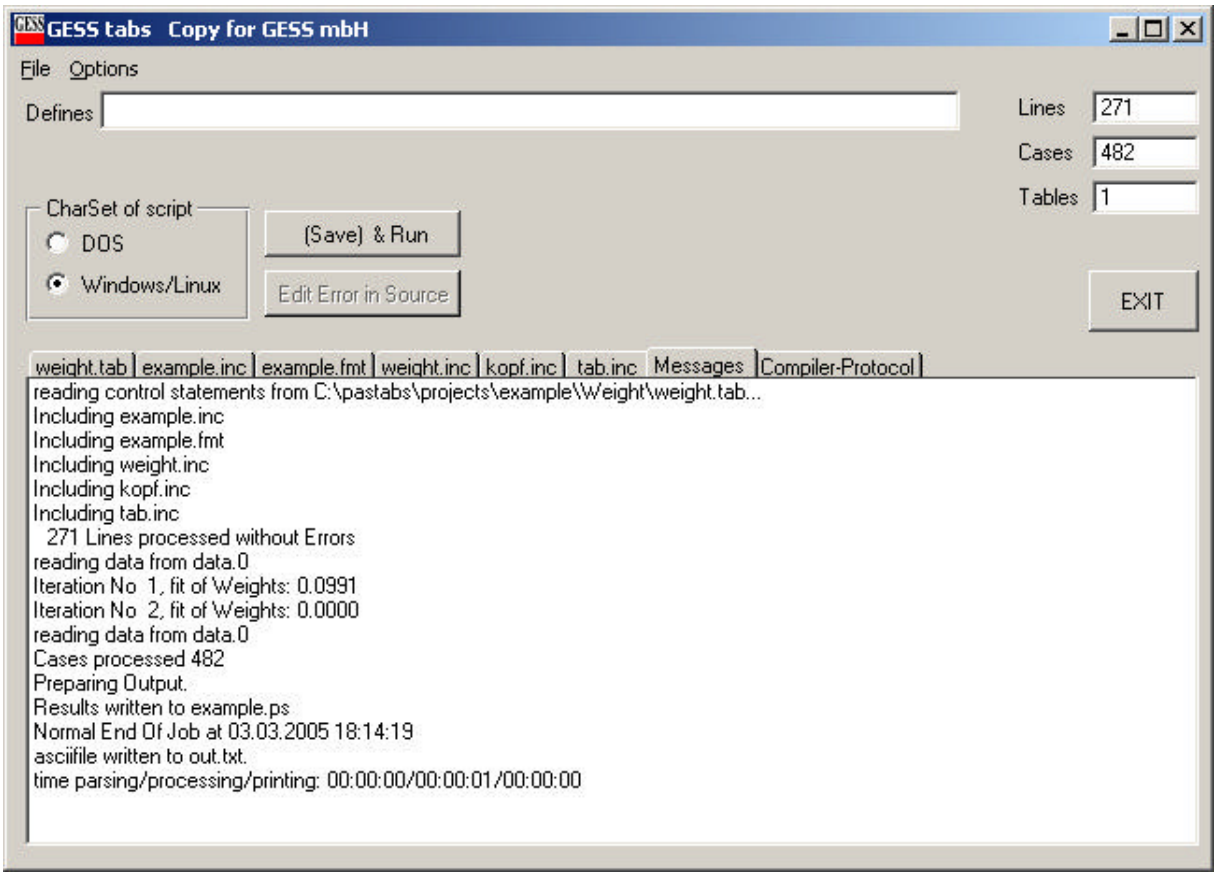
Die Summe aller Vorgaben aller Zellen für eine Variable muss 100% betragen. In WEIGHT.TAB muss jetzt noch die Zeile

```
INCLUDE = weight.inc;
```

ergänzt werden, damit die Gewichtungsvorgabe auch eingesetzt wird.

Wir starten GT.EXE mit dem Anweisungsfile WEIGHT.TAB. In der Lasche „Messages“ sieht man (siehe Abbildung unten), dass GESS tabs im Anschluss an das Einlesen der Daten ein iteratives Verfahren startet, um die erforderlichen Fallgewichte zu ermitteln. In der zweiten Iteration ist exakte Anpassung erreicht (fit of Weights 0.0000). Das ist das normale Verhalten, wenn nur eine Variable angepasst werden muss.

Im Anschluss an die Iterationen wird der Datensatz ein zweites Mal gelesen; erst im zweiten Durchlauf werden die Tabellen gezählt. Der erste Durchlauf dient zur Ermittlung der Zellenbesetzungen als Grundlage der Gewichtung.



Nun kann man bei GESS tabs beliebig[1] viele Variablen zur Gewichtung heranziehen. Für unser Beispiel sollen zusätzlich die Altersgruppen angepasst werden. Das Alter wurde in 3 Ausprägungen erhoben:

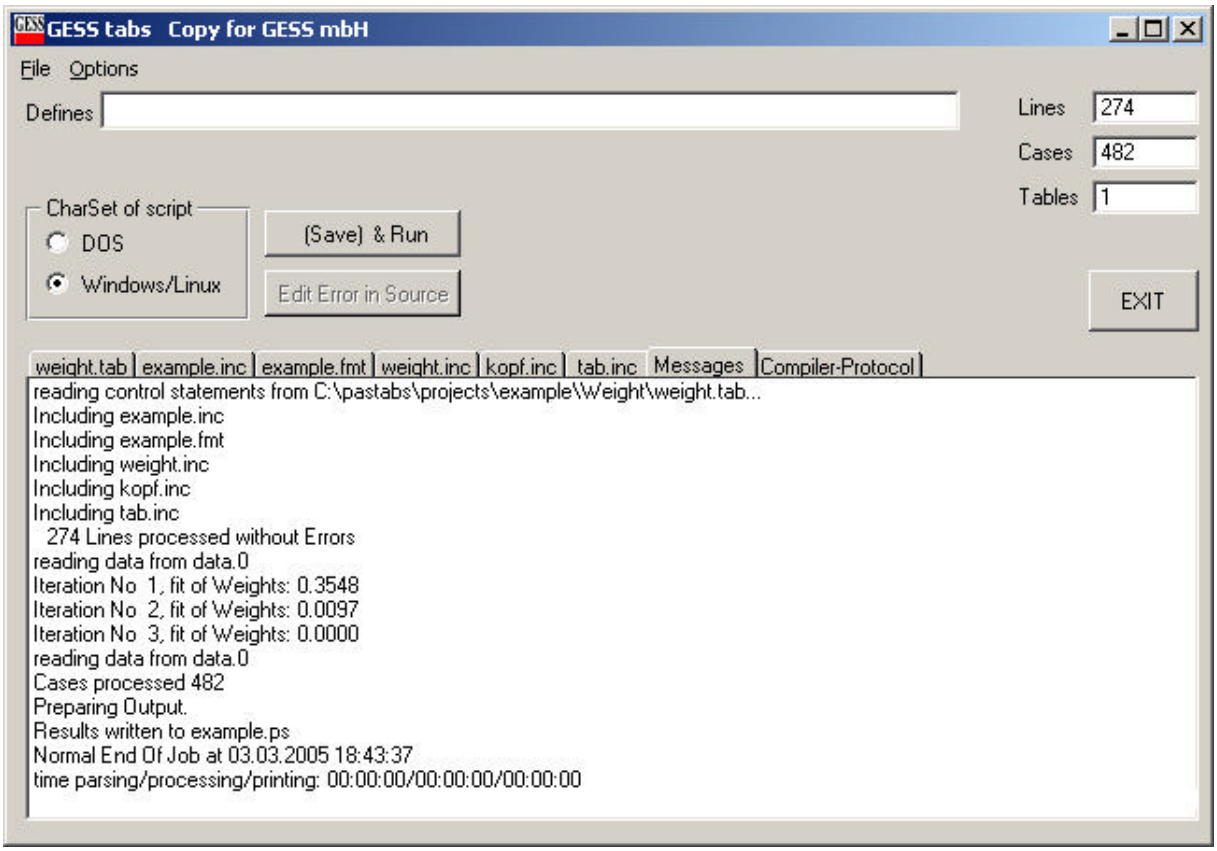
```
variable alter = * 1 labels
1 "18-29 Jahre"
2 "30-49 Jahre"
3 "50-64 Jahre"
;
```

Die Vorgaben seien 24% für die 18-29-jährigen, 28% für die 30-49-jährigen, und 48% für die Befragten von 50-64 Jahren. In WEIGHT.INC wird entsprechend ergänzt:

```
weightcells alter =
1 : 24%
2 : 28%
3 : 48%
;
```

In der folgenden Abbildung sieht man wieder die Messages von GESS tabs:

[1] Das bedeutet nur, dass es kein technisches Limit gibt. In der Praxis erhöht sich parallel zur Anzahl der Gewichtungsvektoren auch das Risiko, dass der Iterationsprozess nicht konvergiert. Man sollte also vorsichtig damit umgehen.



Im Beispiel wurden jetzt 3 Iterationen benötigt. GESS tabs führt bis zu 50 Iterationen durch. Ist bis dahin keine Konvergenz erreicht, wird der Lauf mit einem Fehler abgebrochen. Als Ergebnis sehen wir folgende Tabelle:

Gewichtet:

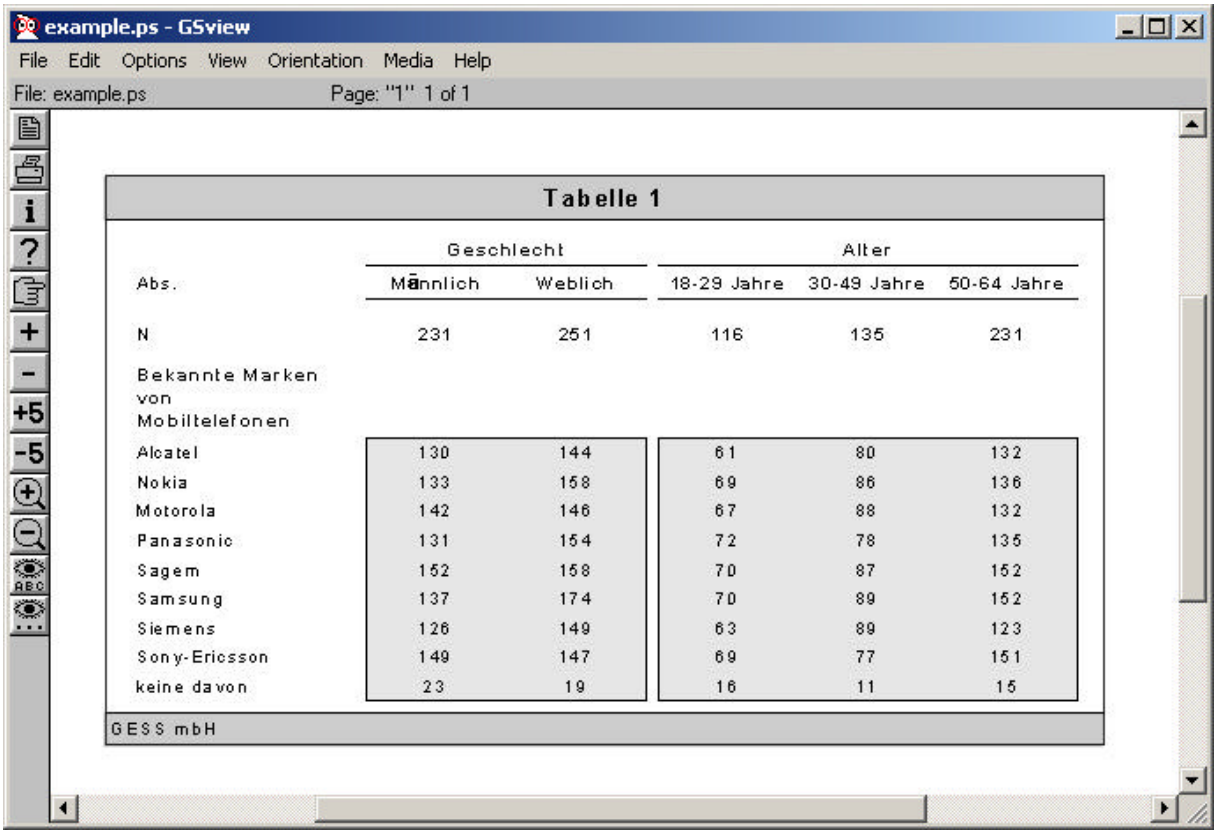


Tabelle 1

Abs.	Geschlecht		Alter		
	Männlich	Weblich	18-29 Jahre	30-49 Jahre	50-64 Jahre
N	231	251	116	135	231
Bekannte Marken von Mobiltelefonen					
Alcatel	130	144	61	80	132
Nokia	133	158	69	86	136
Motorola	142	146	67	88	132
Panasonic	131	154	72	78	135
Sagem	152	158	70	87	152
Samsung	137	174	70	89	152
Siemens	126	149	63	89	123
Sony-Ericsson	149	147	69	77	151
keine davon	23	19	16	11	15

GESS mbH

Wir haben in der Tabelle die absoluten Häufigkeiten ausgeben lassen, so kann man besser die Unterschiede erkennen, die die Gewichtung hervorruft. In der ungewichteten Datei (siehe unten) waren z.B. mehr Männer als Frauen. In der gewichteten Tabellierung (siehe oben) ist dieser Mangel behoben.

Ungewichtet:

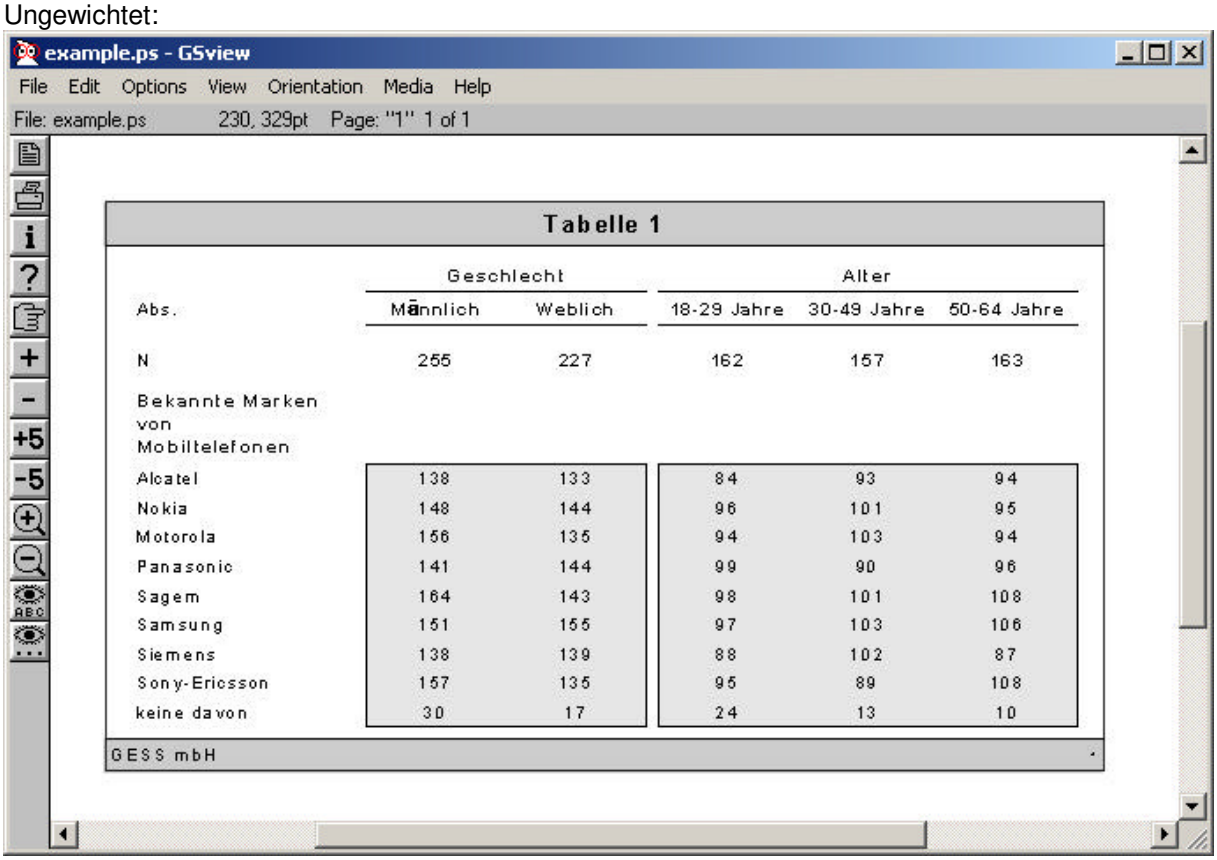


Tabelle 1

Abs.	Geschlecht		Alter		
	Männlich	Weblich	18-29 Jahre	30-49 Jahre	50-64 Jahre
N	255	227	162	157	163
Bekannte Marken von Mobiltelefonen					
Alcatel	138	133	84	93	94
Nokia	148	144	96	101	95
Motorola	156	135	94	103	94
Panasonic	141	144	99	90	96
Sagem	164	143	98	101	108
Samsung	151	155	97	103	106
Siemens	138	139	88	102	87
Sony-Ericsson	157	135	95	89	108
keine davon	30	17	24	13	10

GESS mbH

Nun ist es häufig sinnvoll, die Ergebnisse der Gewichtung (die einzelnen Fallgewichte) im Datensatz zu speichern, um sie z.B. bei multivariaten Analysen außerhalb von GESS tabs zu verwenden.

Die einfachste Methode hierfür ist es, ein COPYFILE erzeugen zu lassen, und über die WEIGHTOUT-Anweisung einen Spaltenbereich für die Ausgabe des Fallgewichts zu vereinbaren:

```
COPYFILE = data.gew;
WEIGHTOUT = 50 8;
```

In der Datei data.gew findet man dann ab Sp. 50 die verwendeten Fallgewichte:

```
     1              2141358MMMM3MMMMMMMM31545      0.79477                          *
     2              23158237MMM9MMMMMMMM11452      1.58919                          *
     3              119MMMMMMMM9MMMMMMMM13521      0.64097                           *
     4              13681423MMM268MMMMMM52332      1.28166                          *
     5              223462158MM9MMMMMMMM53241      0.95366                          *
     6              22327845MMM4352MMMMM24355       0.95366                          *
     7              1314758MMMM178MMMMMM23452      1.28166                          *
     8              22524MMMMMM542MMMMMM54411      0.95366                          *
     9              23745863MMM7564MMMMM23553      1.58919                          *
    10              2365412MMMM9MMMMMMMM25125      1.58919                          *
    11              2356481MMMM9MMMMMMMM21512      1.58919                           *
    12              211465328MM3145MMMMM31115      0.79477                          *
    13              1363458MMMM345MMMMMM54154      1.28166                          *
    14              117234MMMMM73MMMMMMM11544      0.64097                           *
    15              119MMMMMMMM9MMMMMMMM25353      0.64097                          *
    16              224718MMMMM9MMMMMMMM54143      0.95366                          *
    17              23123546MMM9MMMMMMMM31335      1.58919                          *
```

Soweit, so gut. In der Praxis hat man es allerdings in der Regel mit einem weiteren Problem zu tun: „Echte“ Datensätze sind oft nicht so schön vollständig wie die in unserem Beispiel. So hat man es zwar seltener mit fehlenden Angaben zum Geschlecht zu tun, aber schon beim Alter muss man mit einem Anteil von Fällen rechnen, bei denen diese Eigenschaft nicht erhoben werden konnte. Die Gewichtung soll aber natürlich auch dann funktionieren, wenn Variablen mit MISSING Values behaftet sind.

Was wäre denn ein sinnvolles Vorgehen, wenn man nach einer Variablen gewichtet, bei der es Fälle gibt, wo die Ausprägung dieser Variablen nicht bekannt ist? Die übliche Antwort hierauf ist: Der Anteil der Fälle mit MISSING Values soll nach der Gewichtung genau so groß sein wie vor der Gewichtung. Man könnte also die z.B. Variable Alter um ein viertes Label „keine Antwort“ ergänzen, dann müsste man die Variable auszählen, den Anteil der fehlenden Werte ermitteln und die Sollverteilung so umrechnen (stauchen), dass alle Vorgaben zusammen 100% ergeben.

Das wäre alles doch recht umständlich. Hierfür gibt es eine spezielle MISSING Klausel im WEIGHTCELLS Statement, die dies mehr oder weniger automatisch erledigt. Zusätzlich zu den tatsächlich auftretenden Werten macht man eine numerische Vorgabe zu den MISSING Werten. Und die Vorgabe 0% wird hierbei von der Software so interpretiert, dass der empirisch vorgefundene Wert verwendet werden soll. Also, kleine Ergänzung in WEIGHT.INC:

```
weightcells geschl =
1 : 48%
2 : 52%
MISSING : 0%
;

weightcells alter =
1 : 24%
2 : 28%
3 : 48%
MISSING : 0%
;
```

Wenn also im Datensatz beispielsweise für 10% der Fälle das Alter nicht erhoben werden konnte, macht GESS tabs vor dem Beginn der Anpasssungsiterationen folgende Umrechnung:

```
weightcells alter =
1 : 21.6%
2 : 25.2%
3 : 43.2%
MISSING : 10%
;
```

Im Resultat sind also hinterher die Fälle so gewichtet, dass die Submenge der Befragten, die bei Alter eine Antwort gegeben haben, der vorgegebenen Verteilung entsprechen, und alle Befragten, deren Alter nicht bekannt war, haben ein durchschnittliches Gewicht von 1.0.