Prof. Dr. J. Althammer
Dr. U. Pötter
Empirische Gesundheitsforschung

Ruhr-Universität Bochum
Sommersemester 2007

Arbeitsblatt 2

1) Besorgen Sie sich die SHARE Daten für Deutschland von `http://www.stat.rub.de/teaching.html`. Entpacken Sie die Daten in einem Unterverzeichnis. Das Unterverzeichnis entält dann Daten der folgenden „Module“ im Stata Format (*.dta):

Module	Content	Respondents
CV	Coverscreen	All individuals
DN	Demographics	All individuals
PH	Physical Health	All individuals
BR	Behavioural Risks	All individuals
CF	Cognitive Function	All individuals
MH	Mental Health	All individuals
HC	Health Care	All individuals
EP	Employment and Pensions	All individuals
GS	Grip Strength	All individuals
WS	Walking Speed	All individuals
CH	Children	Family respondent
SP	Social Support	Family respondent
FT	Financial Transfers	Financial respondent
HO	Housing	Household respondent
HH	Household Income	Household respondent
CO	Consumption	Household respondent
AS	Assets	Financial respondent
AC	Activities	All individuals
EX	Expectations	All individuals
IV	Interviewer Observations	All individuals
Q	Self-administered Questionnaire	All individuals

sowie einige Dateien mit generierten Variablen über Einkommen, Haushaltszusammensetzung, Schulbildung etc. Ein Codebuch gibt es unter `http://www.share-project.org/new_sites/Documentation/view.pdf`.

`sharerel1_cv_rD.dta` gibt an, wer zu welchen Teilen des Fragebogens Auskunft gegeben hat. Zudem sind einige wichtige Daten wie Alter und Geschlecht enthalten. In `sharerel1_dnD.dta` sind weitere grundlegende demografische Angaben enthalten, `sharerel1_epD.dta` enthält Angaben über Beschäftigung, Einkommen und Pensionen. `sharerel1_hcD.dta` enthält Angaben zur Inanspruchnahme des Gesundheitssystems, zur Krankenversicherung und zu Zuzahlungen etc.

Zunächst sollen die wichtigsten demografischen Daten eingelesen werden. Dazu muss zunächst `R` das Arbeitsverzeichnis mit den Daten angegeben werden (unter: `Datei -> Verzeichnis wechseln`). Dann:

```
library(foreign)  # liest "fremde" Dateiformate
                  # (SPSS, STATA, SAS etc.)
### cover
cv <- read.dta("sharerel1_cv_rD.dta", convert.factors=T)
attach(cv)
```

Die Option `convert.factors=T` liest entsprechende Variable mit ihren Wertebezeichnungen ein. In `R` werden sie dann als Faktoren behandelt. Das ist für den Anfang hilfreich, weil die „Label“ der Werte ausgegeben werden. Auf der anderen Seite ist es meist einfacher, mit Variablen zu arbeiten, die nicht Faktoren sind. Dazu können die Daten mit der Option `convert.factors=F` eingelesen werden.

`cv` ist nun ein „Data Frame“. `dim(cv)` gibt die Anzahl der Fälle und der Variablen an. `names(cv)` gibt die Namen aller Variablen in `cv` an. `summary(cv)` gibt einige deskriptive Statistiken für alle Variablen aus.

Variablen in `cv` können etwa durch `cv$gender` angesprochen werden. Nach `attach(cv)` können die Variablen auch direkt mit ihren Namen (ohne `cv$` davor) angesprochen werden. Aber Achtung: `attach()` macht eine Kopie des „Data Frame“ `cv` und macht sie im „Search Path“ zugänglich. Wenn man also `gender` umkodiert, dann wird `cv$gender` *nicht* geändert! Zudem werden Variablen in `attached` „Data Frames“ in der Reihenfolge gesucht, in der ein `attach()` Kommando angegeben wurde. Man sollte also immer nur ein „Data Frame“ zu jeder Zeit als `attach`ed deklarieren. Ein `attach(cv)` kann durch `detach(cv)` rückgängig gemacht werden. `ls()` gibt eine Liste aller definierten Variablen an, die direkt zugänglich sind. `search()` gibt den gegenwärtigen „Search Path“ an.

Aufgaben: a) Berechnen Sie eine Häufigkeitsverteilung von `gender` im „Data Frame“ `cv`. (`table(cv$gender)`). `gender` ist als Faktor eingelesen worden. Um die Werte „refusal“ und „don’t know“ auszuschliessen, kann `unclass(gender)` benutzt werden.

b) Geben Sie die Altersverteilung (`yrbirth` Geburtsjahr) an. Beachten Sie die Kodierung von `yrbirth`! Definieren Sie eine Variable `age`, die nur valide Altersangaben enthält.

c) Plotten Sie ein Histogramm der Altersverteilung. Benutzen Sie `hist(age)` und probieren Sie einige der Optionen `freq, col, xlab, ylab`.

d) Geben Sie die Altersverteilung getrennt nach Geschlecht an. Der Code könnte etwa wie folgt aussehen:

```
par(mfrow=c(1,2))
hist(age[geschl==0], col="lightblue", main="M\"anner",
     xlab="Alter", freq=F, xlim=c(50,100), ylab="Dichte")
hist(age[geschl==1], col="palevioletred1", main="Frauen",
     xlab="Alter", freq=F, xlim=c(50,100), ylab="Dichte")
par(mfrow=c(1,1))
```

Hier wurde `geschl <- unclass(gender) - 1` definiert.

d) Probieren Sie einige der Alternativen zur Darstellung bedingter Verteilungen, etwa `spineplot(as.factor(geschl) ~ age)` und `cdplot(as.factor(geschl) ~ age )`.

2) Die Variable `sampid2` gibt an, zu welchem Haushalt eine Beobachtung gehört. `cvid` gibt innerhalb eines Haushalts eine Befragtennummer an, zu denen aber auch Personen gehören können, die nicht zur Grundgesamtheit von SHARE gehören. `respid` gibt die Nummern der Befragten im Haushalt an, die auch zur Grundgesamtheit gehören. Eine eindeutige Zuordnung von Daten aus verschiedenen Datensätzen geht also nur durch die gemeinsame Verwendung der beiden Variablen `sampid2` und `cvid`.

Aufgaben: a) Was ist die Anzahl von Haushalten in diesem Datensatz? Was ist die Anzahl von Personen? *Hinweis:* `unique(x)` erzeugt einen Vektor, in dem mehrfach auftretende Elemente in `x` entfernt sind. `length(x)` gibt die Anzahl der Elemente des Vektors `x` an.

b) Geben Sie die Verteilung der einbezogenen Personen je Haushalt an. *Hinweis:* `tapply(x, index, function)` wendet die Funktion `function` nacheinander auf alle Elemente von `x` an, die identische Werte des Vektors `index` enthalten.

```
a <- tapply(sampid2, sampid2, length)
```

liefert also den Vektor, der für alle Haushalte die Anzahl der befragten Personen enthält.

3) Als nächstes soll der Datensatz mit weiteren demografischen Angaben eingelesen werden:

```
### Demografie
dn <- read.dta("sharerel1_dnD.dta", convert.factors=T)
### Zusammenfuehren:
dat <- merge(cv,dn,by=c("sampid2","cvid"))
detach(cv)
attach(dat)
names(dat)
```

Aufgaben: a) Informieren Sie sich in der Hilfe über die verschiedenen Optionen der Funktion `merge()`. Würde `dat <- merge(cv, dn)` zum gleichen Ergebnis führen?

b) Vergleichen Sie das angegebene Geburtsjahr aus dem Fragebogenteil `cv` mit der entsprechenden Angabe im Teil `dn` (Variable `dn003.`). Beachten Sie die Kodierung fehlender Werte. Benutzen Sie auch `summary()`. Wieviele fehlende Werte gibt es für die Variable `dn003.`? Wie gross sind die maximalen Abweichungen zwischen `yrbirth` und `dn003.`?

c) Geben Sie die Verteilung des höchsten Schulabschlusses (`dn010.`) an. Suchen Sie die Bedeutung der Kodierung für Deutschland auf `http://www.share-project.org`.

d) Berechnen Sie die Verteilung des höchsten Schulanschlusses getrennt nach Geschlecht.

4) Nun sollen auch noch die Daten über Beschäftigung sowie die Daten über Gesundheitsnachfrage zusammengeführt werden:

```
### detach!
detach(dat)
### Beschaeftigung
ep <- read.dta("sharerel1_epD.dta", convert.factors=T)
### Zusammenfuehren:
dat <- merge(dat,ep,by=c("sampid2","cvid"))
### hc
hc <- read.dta("sharerel1_hcD.dta", convert.factors=T)
### Zusammenfuehren:
dat <- merge(dat,hc,by=c("sampid2","cvid"))
### Aufraeumen
rm(cv,ep,hc,dn)
attach(dat)
```

Aufgaben: a) Berechnen Sie eine lineare Regression mit der Zahl der Arztbesuche (`hc002.`) als abhängiger Variable und den Kovariablen Alter, Geschlecht, Schulabschluss und Beschäftigungsstatus (`ep005.`). Der Befehl hat die Form `erg <- lm(hc002. ~ age2 + geschl + dn010. + ep005.); summary(erg)`. Beachten Sie aber die Kodierung fehlender Werte insbesondere bei `hc002.` sowie die Kodierung von `ep005.`