Übung 1

Rudolf Dutter & Matthias Templ
Computerstatistik (107.258) WS 2007
10. Oktober 2007

TECHNISCHE UNIVERSITÄT WIEN
VIENNA UNIVERSITY OF TECHNOLOGY

1 Hinweis zu den Übungsdaten

Für diese Übung und einige weitere Übungen werden die **EU-SILC** Daten (**Eu**ropean **S**urvey of **I**ncome and **L**iving **C**onditions) vom Jahr 2004 herangezogen. Die Umfrage EU-SILC dient der Europäischen Kommission vor allem zum Monitoring der *Lissabon 2010* Ziele der Europäischen Union. Diese reichhaltigen Daten sollen Auskunft über den Wohlstand und des sozialen Zusammenhanges der Bevölkerung geben. Für diese Übung steht ein 50-Prozent Sample der österreichischen EU-SILC Daten der Statistik Austria zur Verfügung, wobei aus Gründen des Datenschutzes einige Variablen nicht im Datensatz enthalten sind. Bei Verwendung dieser Daten stimmen sie den Nutzungsbestimmungen (siehe Link) zu.

Jede/r Student/in wird ein eigenes Subset an Daten analysieren, wobei sich dieses Subset durch Auswahl von 3 von 9 Bundesländer beschränkt. Zur Auswahl des Subsets wird Ihre Matrikelnummer als *seed* für den Zufallsgenerator von `R` verwendet. Gehen Sie wie folgt vor:

```
R> set.seed(0123456)
R> d <- sample(1:9, 3)
R> sort(d)
[1] 3 7 8
```

Jene/r Übungsteilnehmer/in mit der Matrikelnummer 0123456 soll nun die Daten für die Bundesländer 3, 7 und 8 analysieren. Für die Selektion des Subsets können Sie folgende Befehlszeilen verwenden.

```
R> load("eusilc.RData")
R> dim(eusilc)
[1] 4626  132
R> eusilc1 <- eusilc[eusilc$bundesld %in% d, ]
R> dim(eusilc1)
[1] 1386 132
```

Verwenden Sie bitte nun diesen kleineren Datensatz für die Übung. Die genaue Beschreibung der Variablen ist im File EU-SILC rev 065_04.pdf zu finden bzw. sind die *eigentlichen* Variablennamen als `comment(eusilc)` gespeichert. Verwenden Sie für diese Übung eine `R`-Version $\geq$ 2.5.0.

2 Übungsbeispiele

1. Kodieren Sie die Variable `P109000` (Rezeptgebührenbefreiung), sodass 1 immer „ja“, 0 immer „nein“ und `NA` immer „keine Antwort“ ist und generieren Sie aus dieser Variable einen Faktor mit Hilfe der Funktion `as.factor()`. Idealerweise verwenden Sie für das Umkodieren die Funktion `ifelse()`, z.B.

   ```
   eusilc1[,"P109000"] <- ifelse(is.na(eusilc1[,"P109000"]), "keine Antwort",
                                 ifelse(euslic1[,"P109000"] == 1, "ja",
                                        ifelse(eusilc1[,"P109000"] == 0, "nein", NA)))
   ```

   Kodieren Sie außerdem die Variable `bundesld` zu einem Faktor mit möglichen Labels `Burgenland` (1), `Kärnten` (2), `Niederösterreich` (3), `Oberösterreich` (4), `Salzburg` (5), `Steiermark` (6), `Tirol` (7), `Vorarlberg` (8), `Wien` (9).

2. Wählen Sie Variable `pek_g` (Brutto-Jahreseinkommen) aus und zeichnen Sie ein Histogramm.

3. Erzeugen Sie Histogramme mit der Funktion `histogram()` und Kerndichteschätzer mit der Funktion `densityplot()` für das Brutto-Jahreseinkommen gesplittet einmal nach `sex` (Geschlecht) und einmal nach `bundesld` (Bundesland). Beide Funktionen sind im R-Packet *lattice* zu finden. Verwenden Sie die

Hilfe zur Funktion `?histogramm` um die Funktion richtig zu verwenden. Tipp: Oft ist es in der Hilfe-Seite hilfreich im Abschnitt *Examples* nachzusehen, wie die Funktion angewendet werden sollte.
Vergleichen Sie die Einkommen nach Geschlecht und nach Bundesland. Kann man z.B. daraus schon auf einen *gender gap* schließen? In welchem Bundesland scheint das Einkommen höher zu sein?

4. Berechnen Sie Schätzungen für Lokation, Streuung, für Variable `pek_g` in Ihren Bundesländern und beschreiben Sie verbal die Datenverteilung. Für die automatische Splittung nach Bundesland verwenden Sie am Besten die Funktion `by()`. Für die korrekte Anwendung der Funktion `by()` verwenden Sie am Besten die Hilfe zur Funktion.

3 Abgabe

Bitte senden Sie die Ausarbeitung in Form eines *pdf*-Files (nicht mehr als 3 Seiten) mit den Resultaten (Outputs plus textliche Kommentare) und Listing des Programmkodes (Funktion) an Ihren jeweiligen Gruppenleiter bis 16. Oktober 2007

r.dutter@tuwien.ac.at
oder
m.templ@tuwien.ac.at

Bitte den Namen des *pdf*-File folgendermaßen:

name_UE_1.pdf

wobei 'name' für den Familiennamen steht.

Empfehlenswert ist es, den Bericht mit „Sweave“ (Verknüpfung von R mit LaTeX) zu erstellen. (Vorlage siehe `http://www.statistik.tuwien.ac.at/public/dutt/vorles/` oder http://www.ci.tuwien.ac.at/~leisch/Sweave/.)