Crashkurs der AG Bioinformatik: Biostatistik mit R

Bei der 15. Jahrestagung der DGKL veranstaltet die Arbeitsgruppe Bioinformatik wieder einen Biostatistik-Workshop. Der Kurs findet am

Donnerstag, 27. September von 14:30 bis 17:30 Uhr

statt. Die Teilnehmerzahl ist auf 25 begrenzt. Kursvoraussetzungen sind:

- WLAN-fähiger Laptop (möglichst Windows, s. u.)
- Installation der kostenlosen Software R und R-Studio
- Durchführung des nachfolgend beschriebenen Software-Tests.

Der Fokus der Veranstaltung liegt auf der Durchführung statistischer Tests und der Ausgabe illustrativer Grafiken für Berichte, Vorträge und Publikationen. Grundlage des Kursprogramms ist ein von der AG Bioinformatik vorgeschlagener Themenkatalog für die Weiterbildung zum Klinischen Chemiker (1).

Installation und Testung von R

Klicken Sie unter *www.r-poject.org* links oben bei *Download* auf CRAN[1] und wählen Sie einen Server in Deutschland aus. Ein direkter Link zur Universität Münster ist zum Beispiel *https://cran.uni-muenster.de/*. Klicken Sie auf *Download R for Windows* und *Install R for the first time*. Übernehmen Sie alle Voreinstellungen.

Installationen unter MacOS oder Linux sind möglich, erfordern aber mehr Computerwissen und stimmen bei der Bedienung nicht immer genau mit dem Kursskriptum überein (nicht empfohlen).

Zum Start des Programms klicken Sie auf das blaue R-Icon, das bei der Installation auf dem Bildschirm abgelegt wurde. Ist dieses Icon nicht vorhanden, dann suchen Sie im Programmordner nach *Rgui.exe.*

Zum Testen geben Sie folgende drei Befehle ein und drücken Sie nach jeder Zeile Return:

```
1:100
boxplot(1:100)
summary(1:100)
```

Sie sollten eine Zahlenreihe von 1 bis 100, eine Boxplot-Grafik und einige statistische Kennzahlen wie etwa Mittelwert und Median erhalten. Erläuterungen dazu werden im Kurs gegeben. Nach dem Test können Sie das Programm wieder schließen und alle Nachfragen bezüglich Speicherung verneinen.

Hinweis: In den letzten Ausgaben von KCM erschien eine Skriptenreihe der AG Bioinformatik (1-4) mit Weiterbildungsinhalten der Biostatistik sowie ausführlichen Anleitungen zum Umgang mit R. Die Durcharbeitung der Lektion 3 wird empfohlen, ist aber nicht Voraussetzung für die Kursteilnahme.

Installation und Testung von R-Studio

Im Kurs nutzen wir R-Studio als komfortable Entwicklungsumgebung für R. Den Download finden Sie unter *www.rstudio.com*. Wählen Sie die lizenzfreie Version für Ihr Betriebssystem aus (vorzugsweise Windows). Übernehmen Sie auch hier alle Voreinstellungen.

Zum Starten klicken Sie auf das Icon von R-Studio. Ist dieses nicht vorhanden, dann suchen Sie im Programmordner nach *rstudio.exe*. Stellen Sie Ansicht so ein wie in Abbildung 1 gezeigt.

[1] Comprehensive R Archive Network

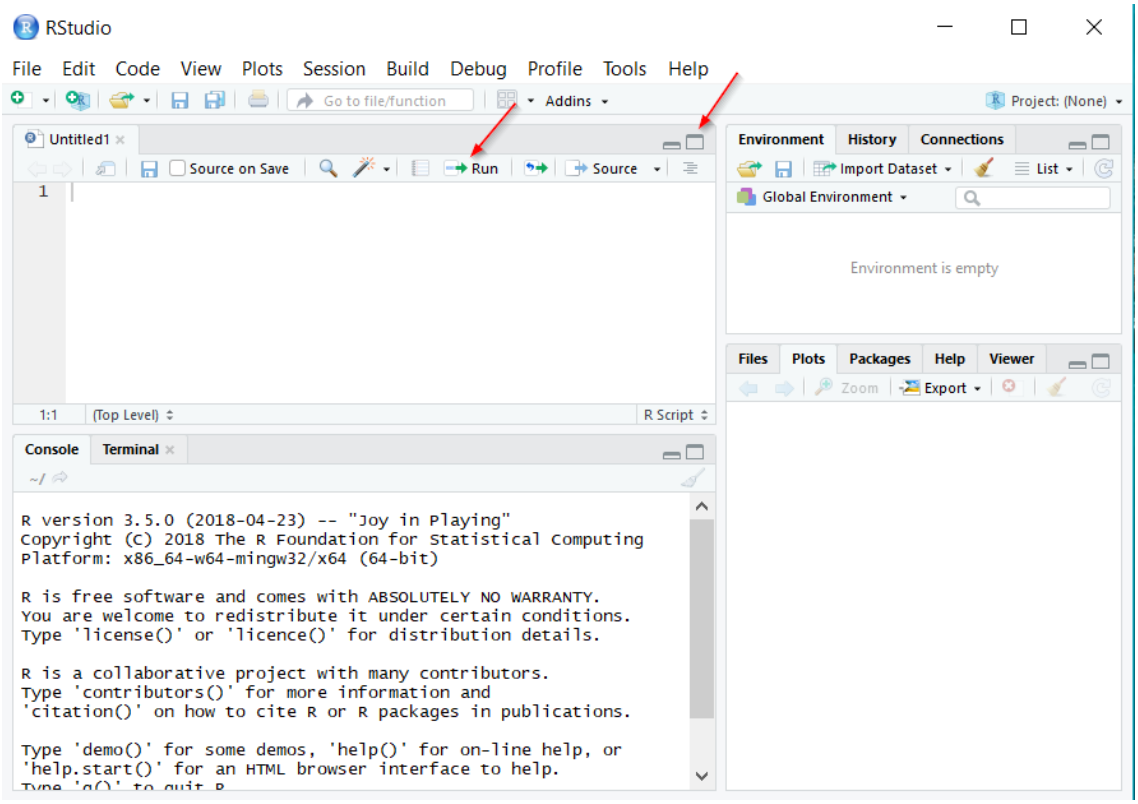

Abb. 1: Die grafische Benutzeroberfläche von R-Studio besteht aus vier Fenstern. Falls beim ersten Öffnen nicht alle Fenster zu sehen sind, klicken Sie auf die Symbole zur Fenstereinstellung jeweils rechts oben. Das Fenster 1 muss gegebenenfalls so breit gezogen werden, dass der Run-Knopf zur Ausführung von R-Programmen zu sehen ist.

Zum Testen tippen Sie links oben erneut *1:100* (ohne Return) ein. Um den Befehl auszuführen, klicken Sie auf *Run*. Sie erhalten im Fenster darunter wieder die Zahlenreihe von 1 bis 100.

Dann geben Sie in die zweite Zeile *box* ein. R-Studio bietet Ihnen daraufhin verschiedene Funktionen an, die mit diesen drei Buchstaben beginnen (Abb. 2). Wählen Sie *boxplot* aus.

Hinter dem Namen fügt R-Studio automatisch die benötigten Klammern ein. Sie müssen nur noch *1:100* eintragen und *Run* anklicken, um die Grafik zu erhalten; Sie erscheint im Fenster rechts unten. Auch bei der dritten Befehlszeile *summary(1:100)* werden Ihnen nach den ersten drei Buchstaben diverse Auswahlmöglichkeiten angezeigt und die Klammern automatisch ergänzt. Die Ausgabe der statistischen Kennzahlen erfolgt im Fenster links unten.

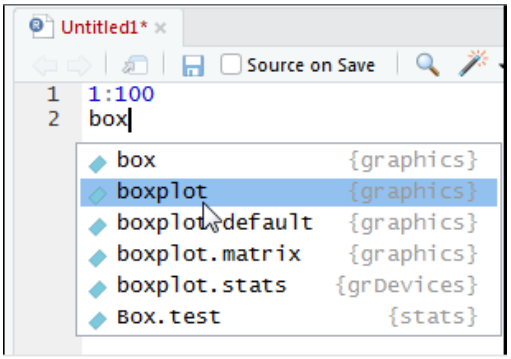

Abb. 2. Wenn Sie in R-Studio die ersten drei Buchstaben eines Funktionsnamens eintippen, erhalten Sie verfügbare Funktionen zur Auswahl.

Als letzten Test zur Vorbereitung des Kurses laden Sie aus dem Internet einen frei verfügbaren Datensatz mit Leberwerten herunter. Geben Sie in Ihre Suchmaschine (z. B. Google) das Suchwort *R datasets* ein und wählen Sie

https://vincentarelbundock.github.io/Rdatasets/datasets.html

Beachten Sie das große R in *Rdatasets*. Sie erhalten eine Liste mit über tausend frei zugänglichen Musterdatensätzen. Suchen Sie in Spalte 1 *texmex* und laden Sie den Datensatz *liver.csv* auf Ihren Rechner.

Kopieren Sie die Datei von *Downloads* in einen geeigneten Ordner und lesen Sie die Daten mit folgender Befehlszeile nach R ein (Zahl der Klammern beachten):

```
x <- read.csv (file.choose())
```

Der Links-Pfeil besteht aus einem Kleiner- und einem Minuszeichen. Nach dem Klick auf *Run* können Sie die Datei mit den Leberwerten auswählen und öffnen.

Wenn alles korrekt funktioniert hat, sehen Sie im R-Studio-Fenster rechts oben, dass Sie 602 Werte (*observations*) der Variablen mit dem Namen x zugewiesen haben. Mit diesem Variablennamen können Sie abschließend ähnliche Funktionen ausführen wie mit der Zahlenreihe von 1 bis 100, z. B.

```
print(x)
boxplot(x)
summary(x)
```

Weitere Erläuterungen folgen im Crashkurs.

Literatur

[1] AG Bioinformatik. Weiterbildungsinhalte aus der Biostatistik und Bioinformatik. KCM 2017; 49(2): 58-60
[2] Hoffmann G. Skriptenreihe der AG Bioinformatik; Grundlage für Kurse und Selbststudium. KCM 2017; 498(3): 114-120
[3] Hoffmann G, Klawonn F. Skriptenreihe der AG Bioinformatik / Lektion 2. KCM 2017; 49(4): 150-159
[4] Hoffmann G, Klawonn F. Skriptenreihe der AG Bioinformatik / Lektion 3. KCM 2018; 50(1): 27-36

Prof. Dr. med. Georg Hoffmann
Medizinischer Fachverlag Trillium GmbH, Grafrath
georg.hoffmann@trillium.de

Prof. Dr. Frank Klawonn, HZI Braunschweig
Dr. Dr. med. Christof Winter, TU München
Dr. Andreas Bietenbeck, TU München