Dynamische Dokumente

Marcus Wurzer

12 Mai 2017

R Markdown

Markdown ist eine einfache Markup-Sprache, mit der HTML-, PDF-, und MS Word-Dokumente erzeugt werden können. `rmarkdown` ist ein Paket, das die Erzeugung solcher Dokumente unter Verwendung von R bzw. RStudio ermöglicht. Diese sogenannten dynamischen Dokumente integrieren R-Code und Fließtext - die Benutzerin/der Benutzer kann beide Komponenten an ihre/seine Bedürfnisse angepasst verflechten ("literate programming").

Da R-Code, -Output, -Grafiken usw. direkt in das Dokument eingebettet sind, ist es sehr einfach, Änderungen vorzunehmen: So ist es z. B. nicht mehr notwendig, eine Grafik in R zu generieren, zu speichern, zu kopieren und in ein Word-Dokument einzufügen - alles geschieht in einem Schritt. Die regelmäßige Erstellung von standardisierten Berichten (z. B. wöchentliche Umsatzberichte) wird dadurch wesentlich erleichtert. Ein weiterer Vorteil, der vor allem im Kontext wissenschaftlichen Arbeitens wichtig ist: Forschung sollte immer reproduzierbar ist, und das wird durch den Literate Programming-Ansatz gewährleistet. Wenn Code und Daten anderen Personen zur Verfügung gestellt werden, so sind diese in der Lage, die Resultate exakt nachzuvollziehen.

Installation

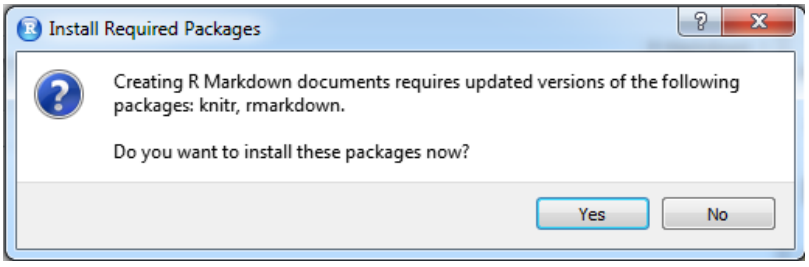
Um R Markdown verwenden zu können, müssen R (Download: https://cran.r-project.org/) und RStudio (Download: https://www.rstudio.com/) installiert sein. R Markdown selbst ist bereits in RStudio inkludiert, aber um PDF-Dokumente erzeugen zu können, müssen wir zusätzlich noch eine TeX-Distribution installieren. (Hinweis: In den EDV-Räumen ist TeX nicht installiert, daher müssen wir uns auf das Erzeugen von Word-Dokumenten beschränken.) Je nach verwendetem Betriebssystem kommen unterschiedliche Distributionen in Frage:

- MiKTeX für Windows-User (Download: http://miktex.org/)
- MacTeX-2016 für OS X-User (Download: https://tug.org/mactex/)
- TeX Live 2016 für Linux-User (Download: https://www.tug.org/texlive/)

RStudio erkennt die installierte TeX-Version dann automatisch.

`rmarkdown` und `knitr`

Wenn TeX erfolgreich installiert wurde, können wir im *File -> New File* -Menü die Option *RMarkdown...* auswählen. Bitte beachten Sie, dass zusätzlich noch die zwei R-Packages (`rmarkdown` und `knitr`) installiert sein müssen. Ist dies nicht der Fall, erscheint die folgende Benachrichtigung:

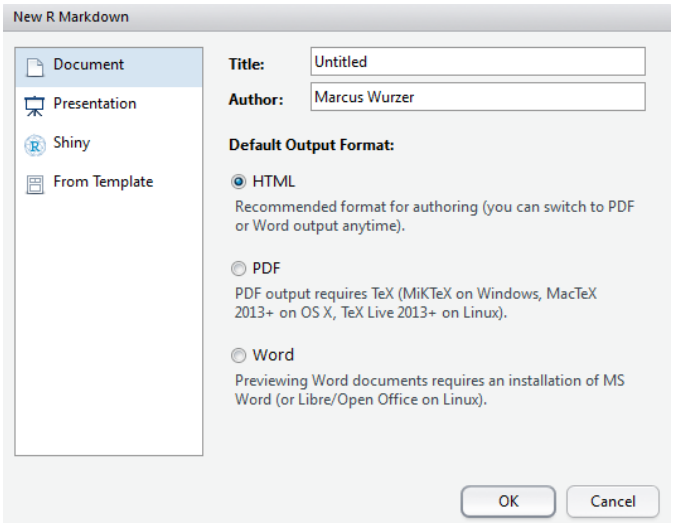


Manchmal funktioniert die Installation über das GUI (d. h., durch Auswahl von *Yes*) nicht richtig. In diesem Fall müssen wir `rmarkdown` manuell über die Konsole installieren (`knitr` wird automatisch mitinstalliert):

```
install.packages("rmarkdown")
```

Erzeugen von Dokumenten

Sobald die beiden genannten Packages erfolgreich installiert wurden, erscheint das folgende Menü:



Hier sollte man seinem Dokument einen Titel geben, das *Author*-Feld ausfüllen und *PDF* als *default output format* setzen. Wenn man dann auf *OK* klickt, wird ein Beispieldokument geöffnet, das die Verwendung von R Markdown illustriert. Man sieht einen *Header*,

```
---
title: "Test document"
author: "Marcus Wurzer"
date: "25 März 2017"
output: html_document
---
```

der verschiedene *key: value*-Paare enthält, Fließtext

```
## R Markdown

This is an R Markdown document. Markdown is a simple for
<http://rmarkdown.rstudio.com>.

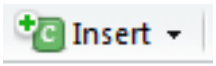
When you click the **Knit** button a document will be ge
document. You can embed an R code chunk like this:
```

inklusive einiger Formatierungsoptionen (die doppelten Sterne auf beiden Seiten des Wortes *knit* sorgen im Output-Dokument beispielsweise dafür, dass **knit** fettgedruckt erscheint). Und dann gibt es noch die R-Chunks:

```
```{r cars}
summary(cars)
```
```

Diese beginnen immer mit drei Backticks und einem Paar geschwungener Klammern (welche den Buchstaben `r` und die Chunk-Optionen beinhalten) und enden wieder mit drei Backticks. Dazwischen befindet sich der

R-Code, welcher ausgeführt wird, sobald die Benutzerin/der Benutzer die Erzeugung des Dokuments startet. Weitere R Code-Chunks können einfach über das Menü *Insert* hinzugefügt werden:



R-Code kann auch innerhalb des Fließtexts verwendet werden. Tippt man z. B. folgende Zeile ins R Markdown-File ein

```
Die Durchschnittsgeschwindigkeit aller Autos im Datensatz beträgt `r mean(cars$speed)` mph.
```

so erhält man im PDF das folgende Ergebnis:

Die Durchschnittsgeschwindigkeit aller Autos im Datensatz beträgt 15.4 mph.

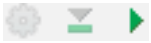
Wenn die Nutzerin/der Nutzer schließlich das Dokument generieren will, klickt sie/er dazu auf den *Knit*-Knopf (Hinweis: Dieser Knopf führt auch in ein Dropdown-Menü, über das man das Output-Format bestimmen kann - PDF, Word, HTML- und zwar unabhängig von der Wahl, die man anfänglich getroffen hat.):

Wenn das R Markdown-File bisher nicht gespeichert wurde, so wird man jetzt durch das Programm dazu aufgefordert. R Markdown-Files haben die Dateiendung .Rmd. Sobald das Dokument erzeugt wurde, öffnet es sich im zugeordneten Programm (Word oder Libre/Open Office für Word-Dokumente, ein PDF Viewer für PDF's, ein Browser oder der in RStudio integrierte Viewer für HTML-Dokumente).

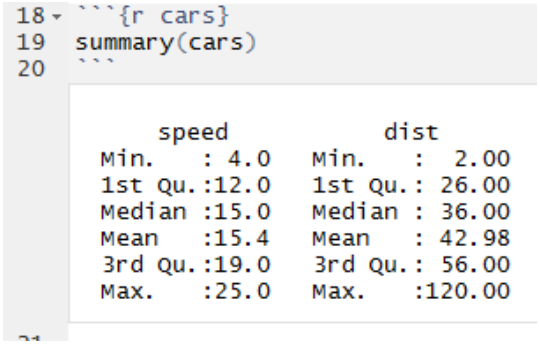
Eventuell braucht MiKTeX noch ein paar Zusatzpakete für die Erzeugung des Dokuments. Wenn ein entsprechender Dialog erscheint, klickt man einfach solange auf *Installieren*, bis alle notwendigen Pakete geladen wurden. Achtung: Das funktioniert nur, wenn man RStudio als Administrator gestartet hat!

R Notebooks

Es kommt relativ häufig vor, dass .Rmd-Files Fehler aufweisen. Versucht man dann, das Dokument zu kompilieren, so wird der Prozess mit einer Fehlermeldung abgebrochen. Diese Meldungen können recht kryptisch sein und sich auf den R-Code oder die Markup-Sprache selbst beziehen. Speziell bei längeren Dokumenten kann die Fehlersuche recht mühsam werden. In diesem Fall ist es eventuell sinnvoll, alle R-Chunks separat auszuführen, um zu sehen, was funktioniert und was nicht. Wir klicken dazu einfach auf den grünen Pfeil, der am rechten Rand jedes Chunks zu finden ist:



Potenzielle Fehlermeldungen und/oder Warnungen werden dann für den jeweiligen Chunk angezeigt. Sollte es hingegen keine Probleme geben, so erhält man direkt im R Markdown-File eine Vorschau auf den Output:



````
18 ```{r cars}
19 summary(cars)
20 ```
````

```
     speed           dist
 Min.   : 4.0   Min.   :  2.00
 1st Qu.:12.0   1st Qu.: 26.00
 Median :15.0   Median : 36.00
 Mean   :15.4   Mean   : 42.98
 3rd Qu.:19.0   3rd Qu.: 56.00
 Max.   :25.0   Max.   :120.00
```

Diese Funktionalität ist seit Version 1.0 von RStudio verfügbar, in der R Notebooks als neue Feature eingeführt wurden. R Notebooks stellen einen speziellen Typ eines R Markdown-Files dar, und es gibt auch einen eigenen *R Notebooks*-Eintrag im *File -> New File*-Menü. Man muss diese Unterscheidung aber nicht unbedingt treffen, siehe dazu auch die Online-Dokumentation (http://rmarkdown.rstudio.com/r_notebooks.html): *"Any R Markdown document can be used as a notebook, and all R Notebooks can be rendered to other R Markdown document types."*

Onlinehilfe

Eine umfangreiche R Markdown-Dokumentation, welche eine kurze Einführung, ein Cheat Sheet usw. enthält, findet man hier:

http://rmarkdown.rstudio.com/

Klicken Sie einfach auf *Get started*.

Beispielhafte Auswertung

Die Auswertung auf den folgenden Seiten wurde mit dem File *Wirtschaftspolitik.Rmd* erstellt. Sie bildet das im Skriptum enthaltene *Musterbeispiel: Wirtschaftspolitik* nach (S. 121f). Hinweis: Die hier inkludierte Version entspricht dem PDF-Dokument, das mit dem .Rmd-File erstellt werden kann. Wird hingegen Word als Output-Format gewählt, sieht das Ergebnis etwas anders - vor allem weniger schön formatiert - aus.

In einem wirtschaftspolitischen Schwellenland stehen Wahlen vor der Tür. Die Hauptaufgabe der neuen Regierung wird es sein, das Land aus der katastrophalen wirtschaftlichen Situation zu führen. Verschiedene wirtschaftliche Ansätze, die Krise zu meistern, werden vorgeschlagen, z.B.: Steuernsenkungen, Steuererhöhungen, Inflationspolitik (kurz gesagt: Geld drucken) oder Deficit spending. Natürlich sind die Kandidaten der Parteien sowohl daran interessiert, was die Anhänger der eigenen Partei bevorzugen, als auch an den Präferenzen der Anhänger anderer Parteien. Eine Zufallsstichprobe von 1000 Wählern wurde über ihr bevorzugtes Rezept zur Behebung der Wirtschaftskrise als auch über ihre Präferenz für eine der drei beherrschenden Parteien befragt.

Unterstützen die Daten die Annahme, dass eine Beziehung zwischen politischer und wirtschaftlicher Präferenz besteht?

Variablen:

- partei - Parteipräferenz
  - 1 = Anhänger von Partei A
  - 2 = Anhänger von Partei B
  - 3 = Anhänger von Partei C
- wipol - Wirtschaftspolitische Präferenz
  - 1 = Steuersenkungen
  - 2 = Steuererhöhungen
  - 3 = Inflationspolitik
  - 4 = Deficit spending
- anzahl - absolute Häufigkeit

Problemanalyse

Welche Variablen kommen vor? *anzahl* ist eine metrische Variable, *partei* und *wipol* sind kategoriale Variablen.

Welche Methode ist angebracht? Wir möchten verschiedene Gruppen (Parteianhängerschaften) bezüglich einer kategorialen Variablen (Option für bestimmte wirtschaftspolitische Maßnahme) vergleichen. Diese Fragestellung kann mit einem χ^2-Test auf Homogenität untersucht werden.

Welche Hypothesen können formuliert werden?
Nullhypothese: die Verteilung der relativen Häufigkeiten für die wirtschaftpolitischen Optionen ist gleich in jeder Parteianhängerschaft. ‚Die Option für bestimmte wirtschaftli- che Maßnahmen ist nicht abhängig von der Parteipräferenz'
Alternativhypothese: die Verteilung der relativen Häufigkeiten für die wirtschaftpoliti- schen Optionen ist nicht in jeder Parteianhängerschaft gleich. ‚Die Option für bestimmte wirtschaftliche Maßnahmen ist abhängig von der Parteipräferenz'
Ein Signifikanzniveau von $\alpha = .05$ wird den Verfahren zugrunde gelegt.

Kurzbericht: Wirtschaftspolitik

Eine Zufallsstichprobe von 1000 Wählern wurde über ihr bevorzugtes Rezept zur Behebung der Wirtschaftskrise als auch über ihre Präferenz für eine der drei beherrschenden Parteien befragt. Verschiedene wirtschaftspolitische Ansätze, die Krise zu meistern, wurden vorgeschlagen, z.B.: Steuernsenkungen, Steuererhöhungen, Inflationspolitik (kurz gesagt: Geld drucken) oder Deficit spending. Die Auftraggeber der Studie sind daran interessiert, welche wirtschaftpolitischen Maßnahmen die Anhänger der verschiedenen Parteien bevorzugen. Die Auszählung von 1000 Befragten nach den zwei Variablen liefert folgende Kreuztabelle, die wir in R mit folgendem Code erstellt haben:

```
##                     partei
## wipol                  A    B    C  Sum
##   Deficit spending    61   90   25  176
##   Inflationspolitik  131   88   31  250
##   Steuererhöhungen    38   67   25  130
##   Steuersenkungen    101  282   61  444
##   Sum                331  527  142 1000
```

```
##                     partei
## wipol                  A    B    C
##   Deficit spending  18.4 17.1 17.6
##   Inflationspolitik 39.6 16.7 21.8
##   Steuererhöhungen  11.5 12.7 17.6
##   Steuersenkungen   30.5 53.5 43.0
```

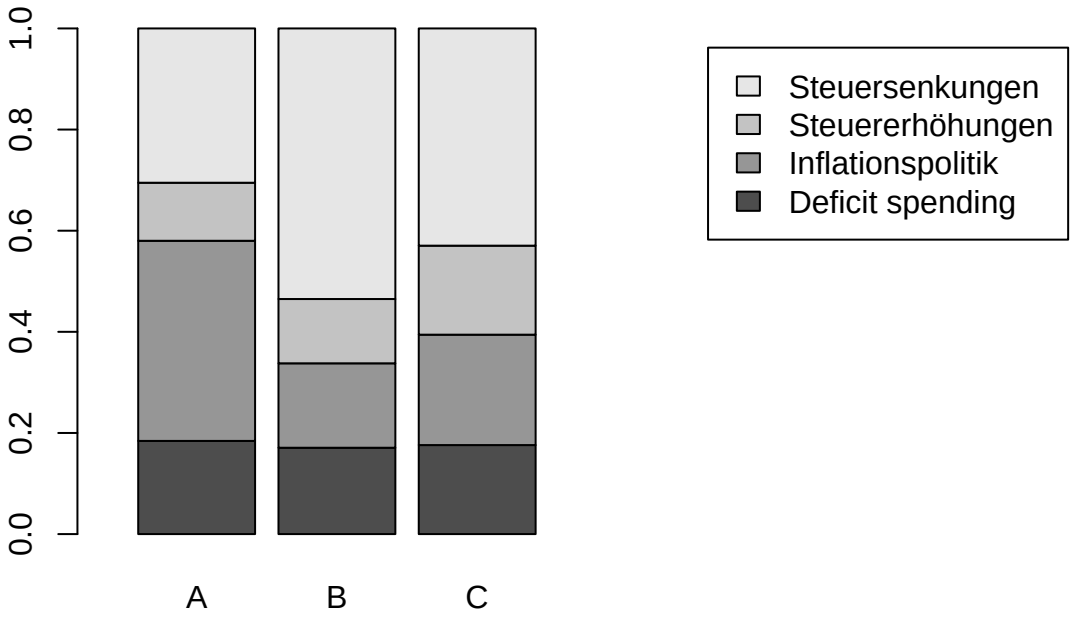
```
## wipol
##  Deficit spending Inflationspolitik  Steuererhöhungen   Steuersenkungen
##              17.6              25.0              13.0              44.4
```

```
## partei
##    A    B    C
## 33.1 52.7 14.2
```

Die Randverteilung der Variable partei zeigt, dass Partei B am meisten Anhänger hat (527, 52.7%), ungefähr ein Drittel unterstützt Partei A (331, 33.1%) und 14.2 Prozent unterstützen Partei C (142).

Weiters sind am meisten für Steuersenkungen (44.4%), ein Viertel ist für Inflationspolitik, 176 Befragte (17.6%) waren für Deficit-spending und 130 (13%) für Steuererhöhungen.

Einen besseren Einblick in die zweidimensionale Tabelle erlangt man über ein Balkendiagramm, bei dem die einzelnen Balken für die drei Parteien stehen.



Den höchsten Prozentsatz an Anhängern einer Steuersenkung gibt es in Partei B. Hingegen ist der Anteil an Anhängern einer Inflationspolitik in Partei A sehr hoch. Der Anteil von Befürwor- tern einer Deficit spending-Politik ist in allen Parteien etwa gleich hoch. Um die drei Parteipräferenzen bezüglich der Wirtschaftspräferenzen zu untersuchen, wird ein χ^2-Test durchgeführt.

```
##
##  Pearson's Chi-squared test
##
## data:  wp
## X-squared = 70.675, df = 6, p-value = 2.973e-13
```

Der Chi-Quadrat Test auf Homogenität bestätigt, dass diese Differenzen sehr groß sind. Der p-Wert ist kleiner als .001 und führt zur Entscheidung, die Nullhypothese zu verwerfen. Man kann also schließen, dass die Präferenzen der Wirtschaftsmaßnahmen in den Anhängerschaften der drei Parteien nicht gleich verteilt sind. Die Wünsche nach bestimmten wirtschaftlichen Maßnahmen sind also nicht in allen Parteianhängerschaften identisch verteilt. Der Wunsch nach Steuersenkungen ist in der Anhängerschaft von Partei B überrepräsentiert, der Wunsch nach Steuersenkungen in Partei A.