Illustrationen von Wahrscheinlichkeiten mit R

Silke Rolles[1]

March 12, 2019

1 Verschiedene Würfel

Mit einem Computer wurden drei Listen mit den Namen `blauerWuerfel`, `roterWuerfel` und `gruenerWuerfel` mit jeweils 60 Zufallszahlen erzeugt:

```
blauerWuerfel

##  [1] 6 4 3 1 1 2 2 5 5 1 6 3 5 1 1 3 6 2 3 2
## [21] 6 6 4 6 3 5 4 3 2 4 2 5 1 1 1 2 4 2 4 6
## [41] 4 6 3 4 1 1 1 2 5 5 2 2 6 2 2 1 2 4 6 3

roterWuerfel

##  [1] 3 5 5 2 6 6 3 6 3 3 6 5 6 4 6 2 3 3 3 6
## [21] 5 6 2 5 3 3 3 5 3 5 6 2 5 6 5 6 3 4 6 2
## [41] 6 5 4 4 2 6 2 6 6 5 6 5 6 6 6 6 6 4 3 6

gruenerWuerfel

##  [1] 4 2 1 3 1 3 5 5 6 3 2 5 6 2 1 4 5 1 6 2
## [21] 2 5 2 5 4 1 1 2 2 6 4 4 3 4 3 2 2 6 4 1
## [41] 4 1 6 5 4 4 4 5 2 4 5 5 3 3 2 2 5 1 2 1
```

Manchmal ist man nicht an den Zufallszahlen selber, sondern an weniger Informationen interessiert, z.B. an den absoluten Häufigkeiten, mit denen die Zahlen $1, \ldots, 6$ auftreten. Hier sind die absolute Häufigkeiten für unsere drei Listen:

```
## blauerWuerfel
##  1  2  3  4  5  6
## 12 14  8  9  7 10

## roterWuerfel
##  2  3  4  5  6
##  7 13  5 12 23
```

[1]Zentrum Mathematik, Bereich M5, Technische Universität München, D-85748 Garching bei München, Germany. e-mail: srolles@ma.tum.de

```
## gruenerWuerfel
##  1  2  3  4  5  6
## 10 14  7 12 11  6
```

Hier sind die zugehörigen relativen Häufigkeiten gerundet auf zwei Nachkommastellen:

```
## blauerWuerfel
##    1    2    3    4    5    6
## 0.20 0.23 0.13 0.15 0.12 0.17

## roterWuerfel
##    2    3    4    5    6
## 0.12 0.22 0.08 0.20 0.38

## gruenerWuerfel
##    1    2    3    4    5    6
## 0.17 0.23 0.12 0.20 0.18 0.10
```

Es fällt auf, dass mit dem roten Würfel keine Eins gewürfelt wurde. Die Wahrscheinlichkeit, dass ein fairer Würfel bei 60 Würfen keine 1 zeigt, beträgt

$$\left(\frac{5}{6}\right)^{60} \approx 1.77 \cdot 10^{-5},$$

was sehr klein ist.
So wurden die Zufallszahlen erzeugt:

```
blauerWuerfel<-sample(1:6,60,replace=TRUE)
```

Der blaue Würfel ist ein fairer Würfel, bei dem alle Zahlen gleich wahrscheinlich sind.

```
roterWuerfel<-sample(c(1,2,3,4,5,6),60,replace=TRUE,
                     prob=c(0,1/6,1/6,1/6,1/6,2/6))
```

Beim roten Würfel tritt die Zahl i mit Wahrscheinlichkeit p_i auf, wobei

$$p_1 = 0, \quad p_2 = p_3 = p_4 = p_5 = \frac{1}{6} \approx 0.167, \quad p_6 = \frac{2}{6} = \frac{1}{3}.$$

D.h. bei einem fairen Würfel wurde die 1 mit einer 6 überklebt.

```
gruenerWuerfel<-sample(c(1,2,3,4,5,6),60,replace=TRUE,
                       prob=c(5/35,6/35,6/35,6/35,6/35,6/35))
```

Beim grünen Würfel tritt die Zahl i mit Wahrscheinlichkeit q_i auf, wobei

$$q_1 = \frac{5}{35} = \frac{1}{7} \approx 0.143, \quad q_2 = q_3 = q_4 = q_5 = q_6 = \frac{6}{35} \approx 0.171.$$

Das ist *kein* fairer Würfel. Die Zahl 1 tritt mit geringerer Wahrscheinlichkeit auf als beim fairen Würfel, während die Zahlen 2,3,4,5,6 gleichwahrscheinlich sind und eine geringfügig höhere Wahrscheinlichkeit haben als beim fairen Würfel.
Das Beispiel zeigt, dass man bei 60 Würfen nur wenig Information über die zugrundeliegenden Wahrscheinlichkeiten gewinnen kann. Insbesondere kann man nicht entscheiden, ob ein Würfel fair ist, es sei denn, er ist ganz deutlich unfair, wie der rote Würfel. Das Beispiel zeigt auch, dass relative Häufigkeiten nicht nahe an den zugrundeliegenden Wahrscheinlichkeiten sein müssen.

2 Erste Befehle mit R

Vorbereitungen. R ist eine freie Programmiersprache für statistische Berechnungen und Grafiken. Die offizielle R Homepage, auf der unter anderem eine ausführliche Beschreibung dieser Sprache angeboten wird, finden Sie unter

`http://www.R-project.org`.

RStudio ist eine Entwicklungsumgebung für R, die frei erhältlich ist unter

`https://www.rstudio.com`.

1. Installieren Sie eine LaTeX-Distribution, zum Beispiel MacTeX (Mac OS) oder MiKTeX (Windows). Mit einem LaTeX-Editor Ihrer Wahl können Sie danach insbesondere LaTeX-Dokumente erstellen.

2. Laden Sie die aktuelle Version von R von `https://cran.rstudio.com` herunter.

3. Installieren Sie nun RStudio auf Ihrem Computer. Befolgen Sie dazu bitte die Anleitungen auf der oben genannten Homepage von RStudio.

4. Zum Erstellen von Handouts, die Sie in RStudio bearbeitet haben, verwenden Sie R Markdown. Installieren Sie R Markdown, indem Sie den Befehl

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

   in die Kommandozeile von RStudio eingeben und ausführen.

Ein erstes Handout erstellen. Wählen Sie in RStudio `File` → `New File` und im Anschluss `R Markdown`. Geben Sie den Titel sowie Ihren Namen ein und wählen Sie als Ausgabeformat PDF. `File` → `Knit` wird nun ein Testdokument erstellt, in welchem Sie weitere Erklärungen zum Umgang mit R Markdown finden.

Erste verwendete R-Befehle. Der Befehl

```
sample(1:6,60,replace=TRUE)
```

erzeugt 60 Zufallszahlen, die durch ziehen aus der Menge $\{1, \ldots, 6\}$ mit zurücklegen (`replace=TRUE`) gezogen werden. Mit dem Befehl

```
blauerWuerfel<-sample(1:6,60,replace=TRUE)
```

werden die Zufallszahlen der Variablen `blauerWuerfel` zugeordnet. Der Befehl

```
sample(c(1,2,3,4,5,6),60,replace=TRUE,
                       prob=c(5/35,6/35,6/35,6/35,6/35,6/35))
```

erzeugt 60 Zufallszahlen, die durch ziehen aus der Menge $\{1, \ldots, 6\}$ gezogen werden, wobei die Zahl $i \in \{1, \ldots, 6\}$ mit Wahrscheinlichkeit q_i gezogen wird und der Vektor $(q_1, \ldots, q_6)$ durch `prob` gegeben ist. Mit dem Befehl

```
blauerWuerfel
```

werden die Werte der Liste `blauerWuerfel` ausgegeben. Der Befehl

```
table(blauerWuerfel)
```

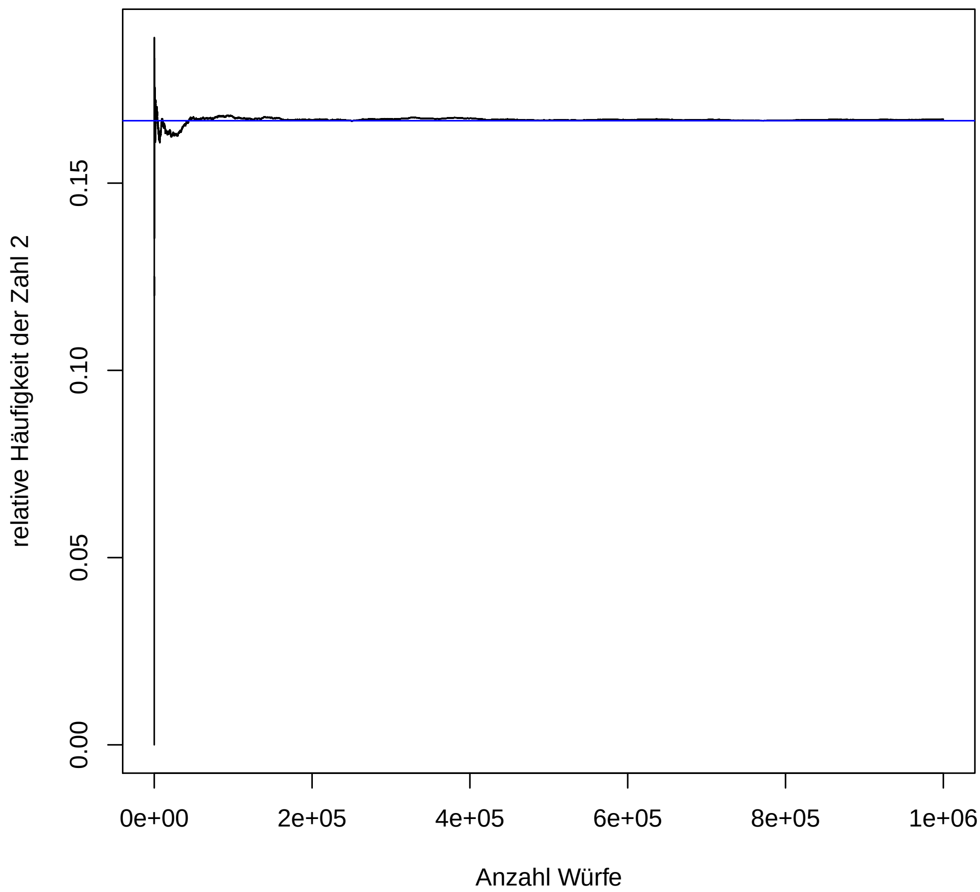
berechnet die absoluten Häufigkeiten in der Liste `blauerWuerfel`. Mit

```
round(table(blauerWuerfel)/60,digits = 2)
```

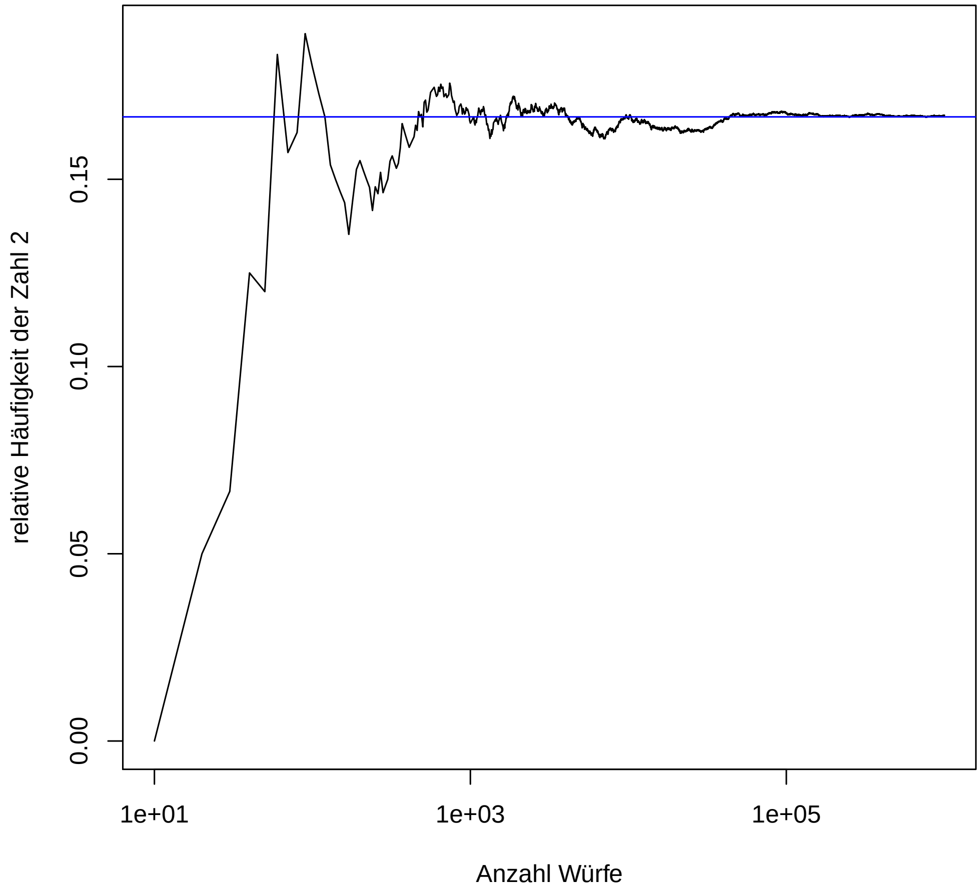
werden die relativen Häufigkeiten berechnet und auf 2 Nachkommastellen gerundet (`digits=2`).

3 Relative Häufigkeiten im Verlauf der Zeit

Was ist die Wahrscheinlichkeit des Ereignisses $\{2\}$ beim einmaligen Wurf eines Würfels? Intuitiv würde man wie folgt vorgehen: Man wiederholt das Zufallsexperiment "Wurf eines Würfels" sehr oft und schaut sich die relativen Häufigkeiten an.
Mit einem Computer wurde eine Million fairer Würfelwürfe simuliert und die relative Häufigkeit der Zahl 2 dargestellt. Die blaue Linie wird durch $y = \frac{1}{6}$ beschrieben.



Die Grafik ist nicht sehr aussagekräftig, da die meisten Fluktuationen bei den ersten Würfen sind. Daher ist es besser, auf der x-Achse eine logarithmische Skala zu verwenden.



Hier sieht man zunächst starke Fluktuationen. Dann nähert sich die relative Häufigkeit dem theoretischen Wert $\frac{1}{6}$ an.
Das Gesetz der großen Zahlen, das wir später beweisen werden, zeigt, dass für jede "typische" Folge $\omega = (\omega_i)_{i\in\mathbb{N}}$ von Würfelwürfen gilt:

$$\text{Wahrscheinlichkeit, eine 2 zu werfen } = \lim_{n\to\infty} h_n(\omega_1, \dots, \omega_n, 2).$$

Hierzu einige R-Befehle. Wir wollen 10^6 faire Würfelwürfe simulieren. Dazu simulieren wir `n` Blöcke mit `k` Würfen. Der Vektor `a` wird für die absoluten Häufigkeiten der Zweien benötigt. Zunächst initialisieren wir alle Variablen. Insbesondere ist `a` ein Vektor der Länge `n`, der mit Nullen initialisiert wird:

```
k=10
n=1e5
a=rep(0,n)
```

In `wuerfel` speichern wir die Ergebnisse von `k` fairen Würfelwürfen. Der Befehl `wuerfel==2` erzeugt einen Vektor, der gleichen Länge wie `wuerfel`, in dem jede 2 durch den Eintrag `TRUE` und jedes andere Ergebnis durch den Eintrag `FALSE` ersetzt wird. `sum(zweien)` berechnet die Summe der Einträge des Vektors `zweien`, wobei `TRUE`=1 und `FALSE`=0; dies liefert also die absolute Häufigkeit der Zweien in der Zufallsfolge `wuerfel`.
Der Vektor `a` besitzt die Einträge `a[i]`, $1 \leq i \leq n$. Es ist `a[1]`=0 und `a[i]` wird die absolute Häufigkeit der Zweien bei den ersten $k(i-1)$ Würfen zugeordnet. Dies wird durch die folgende `for`-Schleife realisiert.

```
for(i in 2:n+1)
{
wuerfel<-sample(1:6,k,replace = TRUE)
zweien<-(wuerfel==2)
a[i]<-a[i-1]+sum(zweien)
}
```

Jetzt wird eine Grafik der relativen Häufigkeiten erzeugt. Der Befehl

```
t<-seq(0,n*k,by=k)
```

erzeugt den Vektor `t` mit den Zahlen $(0, k, 2k, 3k, \dots, nk)$. Für $t = ik$ ist `a[(t/k)+1]/t`=`a[i+1]/(ik)` die relative Häufigkeit der Zweien bei den ersten ki Würfen. In der Grafik wird die Funktion `t`$\mapsto$`a[(t/k)+1]/t` gezeichnet. Mit `xlab` bzw. `ylab` beschriftet man die x- bzw. y-Achse.

```
plot(t,a[(t/k)+1]/t,type="l",xlab="Anzahl Würfe",
     ylab="relative Häufigkeit der Zahl 2")
```

Um die x-Achse auf logarithmischer Skala darzustellen, verwendet man im `plot`-Befehl den Zusatz `log="x"`. Mit dem Befehl

```
abline(h=1/6,col="blue")
```

zeichnen wir eine blaue horizontale Linie by $y = \frac{1}{6}$ ein.

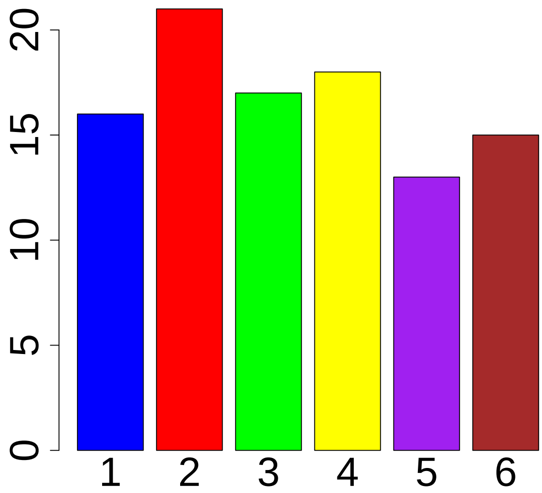
Absolute Häufigkeiten im Verlauf der Zeit. Als nächste möchten wir uns die Histogramme von `wuerfel` im Verlauf der Zeit anschauen. Zunächst simulieren wir $n = 10^5$ faire Würfelwürfe.

```
k=100
n=1e5
wuerfel<-sample(1:6,n,replace = TRUE)
```

Mit dem Befehl

```
barplot(table(wuerfel[c(1:100)]),width=1,
    col=c("blue","red","green","yellow","purple","brown"),
    cex.axis=2.8,cex.names=2.8)
```

erstellen wir ein Histogramm der absoluten Häufigkeiten der ersten 100 Würfe. `c(1:100)` erstellt einen Vektor mit den Zahlen $(1, 2, \ldots, 100)$. Der Befehl `wuerfel[c(1:100)]` schränkt den Vektor `wuerfel` auf die ersten 100 Einträge ein. `table` erstellt eine Tabelle mit den absoluten Häufigkeiten. `barplot` erzeugt das Histogramm.



Mit Hilfe einer `for`-Schleife wird für jedes `i` zwischen `1` und `n/k` ein Histogramm der ersten ki Würfelwürfe erstellt.

```
for(i in 1:(n/k))
{
barplot(table(wuerfel[c(1:(k*i))]),width=1,
    col=c("blue","red","green","yellow","purple","brown"),
    cex.axis=2.8,cex.names=2.8)
Sys.sleep(1)
}
```

`wuerfel`$[$`c`$(1:($`k`$*$`i`$))]$ betrachtet nur die ersten ki Würfe. Die y-Achse wird automatisch angepasst. Fügt man zu `barplot` noch `ylim` $=$ `c`$(0, 200)$ hinzu, so wird für alle Histogramme eine y-Achse von 0 bis 200 gewählt. `Sys.sleep`(1) macht eine Pause von einer Sekunde.