Uebung Nr. 4

Valentina Cisternas Seeger und Roman Kessler

05.11.2019

Uebung 4

Explorative Datenanalyse und Visualisierung

Wintersemester 2019

S. Doehler, B. Nedic (FBMN, h_da)

Name: Roman Kessler und Valentina Cisternas Seeger

Punkte:

Aufgabe 13. Verwenden Sie den R-Datensatz `faithful` (dieser Datensatz ist im base package enthalten). Das Merkmal `eruptions` enthaelt die Laenge von Eruptionen (in Minuten) des Old Faithful Geysirs im Yellowstone National Park in Wyoming, USA. Sie sollen die Verteilung der Eruptionen analysieren. Achten Sie bei allen Grafiken darauf, dass sowohl der Titel der Grafik als auch die Beschriftungen der Achsen moeglichst aussagekraeftig sind.

(a) Laden Sie die Daten und erstellen Sie eine five-point summary.

(b) Erstellen Sie einen Box-Plot und interpretieren Sie diesen.

(c) Erstellen Sie einen Stem and leaf-Plot und interpretieren Sie diesen.

(d) Plotten Sie die empirische Verteilungsfunktion.

(e) Plotten Sie das Histogramm (inklusive eines "rug-plots") und interpretieren Sie dieses. Experimentieren Sie mit verschiedenen Klassenbreiten. Wie muessen Sie die Optionen einstellen, damit das Histogramm als Dichteschaetzer geplottet wird? Wie erhaelt man ein aehnliches Bild wie beim stem and leaf-Plot?

(f) Plotten Sie ein gleitendes Histogramm in das Histogramm mit ein. Experimentieren Sie mit verschiedenen Bandbreiten. Welche Bandbreite liefert ein "gutes" Ergebnis?

(g) Wiederholen Sie Teil (f) mit dem Gauss- und Epanechnikov-Kernen.

(h) Vergleichen Sie die verschiedenen Methoden. Was haben Sie ueber die Daten gelernt? Welche Methode(n) wuerden Sie einem Anwender empfehlen?

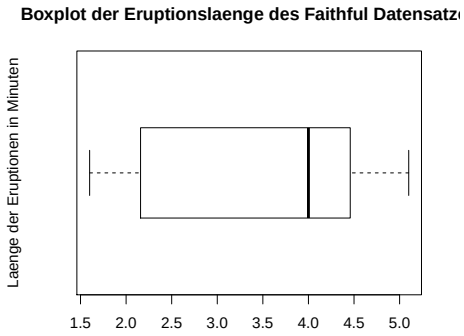
Loesung

Zu (a):

```
summary(faithful)
#>    eruptions        waiting
#>  Min.   :1.600   Min.   :43.0
#>  1st Qu.:2.163   1st Qu.:58.0
#>  Median :4.000   Median :76.0
#>  Mean   :3.488   Mean   :70.9
#>  3rd Qu.:4.454   3rd Qu.:82.0
#>  Max.   :5.100   Max.   :96.0
```

Zu (b):

```
d=faithful$eruptions
boxplot(d,main="Boxplot der Eruptionslaenge des Faithful Datensatzes",
          ylab="Laenge der Eruptionen in Minuten",horizontal = TRUE)
```



Der Median der Eruptionen liegt bei 4 Minuten. Das Minimum liegt ist bei 1.6 Minuten und das Maximum bei 5.1 Minuten. Dies laesst auf eine schiefe Verteilung der Daten schliessen.

Zu (c):

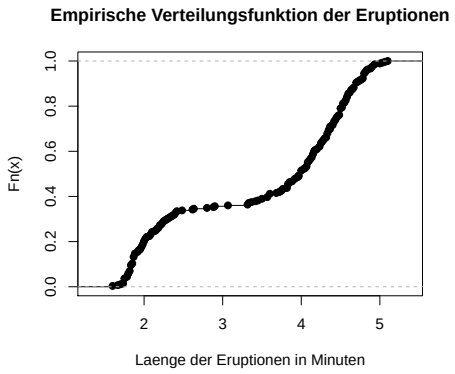
```
stem(d)
#>
#>   The decimal point is 1 digit(s) to the left of the |
#>
#>   16 | 070355555588
#>   18 | 000022233333335577777777788882233577788888
#>   20 | 000022233788000357778
#>   22 | 00023355780235778
#>   24 | 00228
#>   26 | 23
```

```
#>   28 | 080
#>   30 | 7
#>   32 | 2337
#>   34 | 250077
#>   36 | 0000823577
#>   38 | 2333335582225577
#>   40 | 000000335778888800223355557777778
#>   42 | 03335555778800233333555777778
#>   44 | 022223355577800000000233333577778888
#>   46 | 0000233357700000023578
#>   48 | 00000022335800333
#>   50 | 0370
```

Das Stem-and-Leaf Plot gibt schon eine genauere Verteilung der Daten wieder. Wir sehen eine gewisse Dichotomie.Die meisten Eruptionen dauern zw. 1.6-2.2 Minuten und 4.-4.8 Minuten.Es gab nur eine Eruption die 3 Minuten dauert. Zwischen 2.4 und 3.8 ist es eher duennbesetzt.

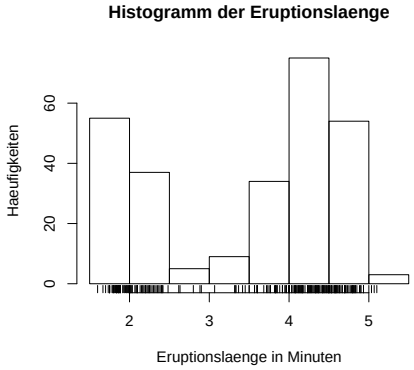
Zu (d):

```
plot(ecdf(d),xlab="Laenge der Eruptionen in Minuten", main="Empirische Verteilungsfunktion d
```

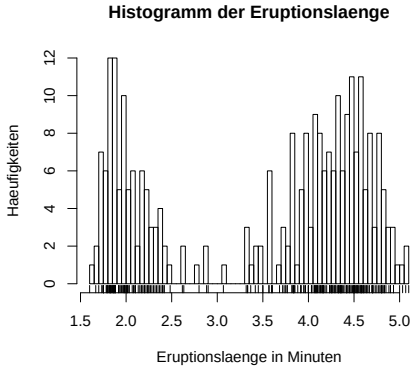


Zu (e):

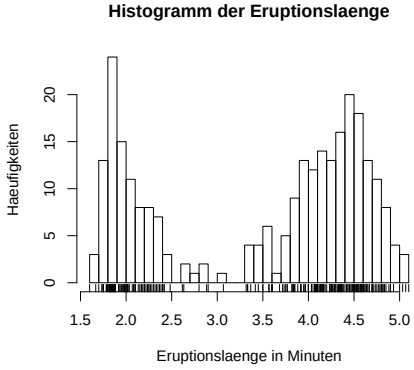
```
hist(d, main="Histogramm der Eruptionslaenge", xlab="Eruptionslaenge in Minuten",ylab="Haeuf
rug(d)
```




```
hist(d, main="Histogramm der Eruptionslaenge", xlab="Eruptionslaenge in Minuten",ylab="Hae
rug(d)
```



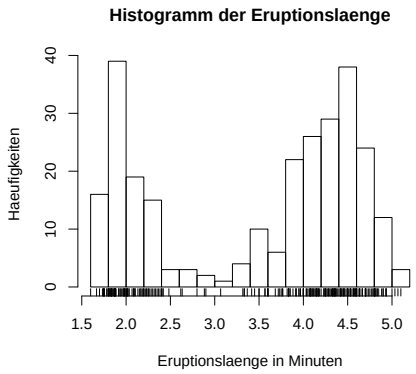

```
hist(d, main="Histogramm der Eruptionslaenge", xlab="Eruptionslaenge in Minuten",ylab="Ha
rug(d)
```



DAs Histogramm (mit dem Rugplot) bestaetigt noch einmal die Vermutung ueber die dichotome Verteilung der Daten. Am haeufigsten dauern die Eruptionen 1.8 Minuten und 4.4 Minuten. Es gibt kaum Eruptionen der Laenge von 2.4 und 3.3. Es gibt ein Paar Ausreisser die in dem Bereich 2.7 bis 3.2 Minuten vorfallen. Die Daten sind vor allem links und rechts angehaeuft.Vor allem die grossen Haeufungen stechen durch den Rugplot hervor.

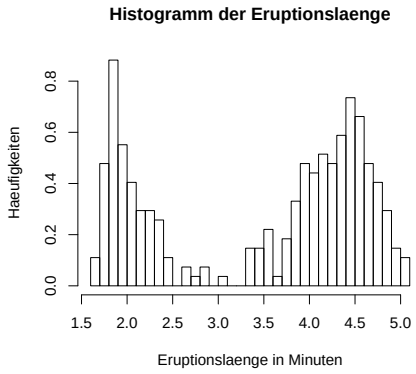
Ein aehnliches Bild wie bei Stem and Leaf Plot wird ilustriert, wenn man das Histogram in 18 Teilen. Dann ist die Bandbreite 0.2.

```
hist(d, main="Histogramm der Eruptionslaenge", xlab="Eruptionslaenge in Minuten",ylab="Haeu
  rug(d)
```



Damit das Histogramm als Dichteschaetzer eingesetzt werden kann, muessen die Summen der Balkenflaechen eins ergeben. Beim Verwenden der relativen Haeufigkeit erhalten wir dies.

```
hist(d, main="Histogramm der Eruptionslaenge", xlab="Eruptionslaenge in Minuten",ylab="Haeuf
```

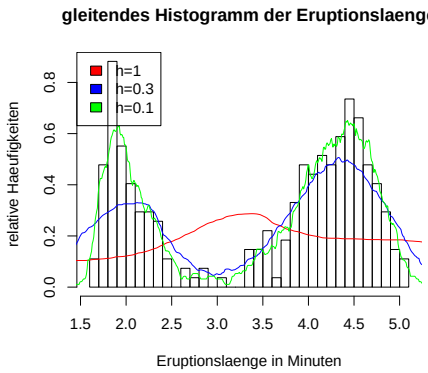


Nun haben wir die relativen Haeufigkeiten verwendet,demenstsprechend ist dann

die Summe aller Flaechen insgesamt 1.

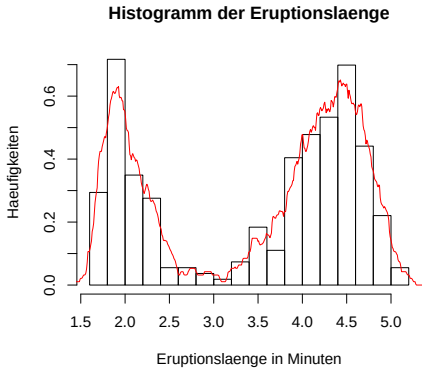
Zu (f): #d=5-1.6=3.4 #damit brauen wir 34 breaks fuer die Bandbreite von 0.1

```
 {hist(d, main="gleitendes Histogramm der Eruptionslaenge", xlab="Eruptionslaenge in Minuten
lines(density(d, kernel='rectangular', 1.0), col='red')
lines(density(d, kernel='rectangular', 0.3), col='blue')
lines(density(d, kernel='rectangular', 0.1), col='green')
 }
legend("topleft",
       c("h=1","h=0.3", "h=0.1"),
       fill=c("red","blue", "green"))
```



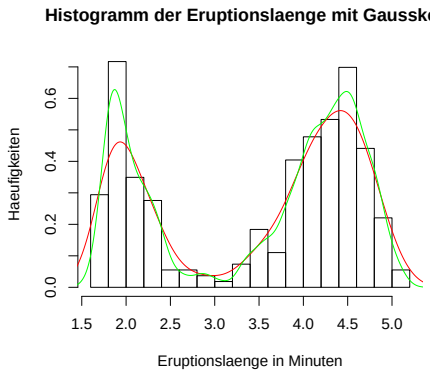
Besonders das gleitende Histogramm mit Bandbreite 0.1 naehert sich gut an das Histogramm an.

```
{
hist(d, main="Histogramm der Eruptionslaenge", xlab="Eruptionslaenge in Minuten",ylab="Haeuf
lines(density(d,kernel="rectangular",0.1), col="red")
}
```

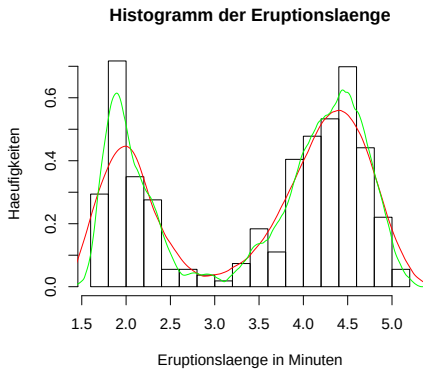


Zu (g):

```
#Gauss Kern
{
hist(d,main="Histogramm der Eruptionslaenge mit Gausskern", xlab="Eruptionslaenge in Minuten
lines(density(d,kernel="gauss",0.2), col="red")
lines(density(d,kernel="gauss",0.1), col="green")
}
```

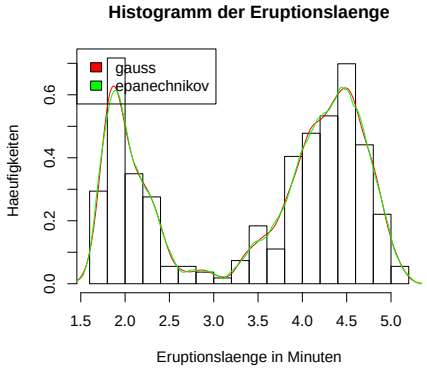



```
#Epanechnikov Kern
hist(d, main="Histogramm der Eruptionslaenge", xlab="Eruptionslaenge in Minuten",ylab="Haeuf
lines(density(d,kernel="epanechnikov",0.2), col="red")
lines(density(d,kernel="epanechnikov",0.1), col="green")
```




```
#Bessere Variante mit Bandbreite 0.1. Da das Maximum besser angenaehert wird.
```

```
hist(d, main="Histogramm der Eruptionslaenge", xlab="Eruptionslaenge in Minuten",ylab="Haeuf
lines(density(d,kernel="gauss",0.1), col="red")
lines(density(d,kernel="epanechnikov",0.1), col="green")
legend("topleft",c("gauss","epanechnikov"),fill=c("red","green"))
```



Der Gauss und auch der Epanechnikov Kernel naehert sich gut der Verteilung an. Verglichen zu dem Graph in (f) sind die Funktionen sehr glatt.

Zu (h): Alles in allem stellt sich heraus, das der Boxplot keine gute Wahl ist die Daten zu plotten, da man nicht sehr gut in der Verteilung sehen kann, dass es zwei Erhoehungen gibt (Dichotomie). Die empirische Verteilungsfunktion war in der Hinsicht schon genauer. Dennoch konnte man nicht gut erkennen, welche Eruptionslaengen am haeufigsten vorkamen. Dies konnte man widerum besser durch das Histogramm sehen, je nach dem welche Schrittbreite gewaehlt wurde. Zu kleine Schrittbreiten hat die Verteilung zu zackig und zittrig werden lassen. Grosse Schrittbereiten widerum fuehrten dazu, das die Haeufigkeitsverteilung sehr ungenau wurde. Durch das glatte Histogramm und vor allem durch den Kernel konnte man eine gute Annaeherung an das Histogramm erreichen mit Bandbreite 0.1.

Anmerkungen/Korrektur
