?bung 4

Explorative Datenanalyse und Visualisierung

Wintersemester 2019

S. D?hler, B. Nedic (FBMN, h_da)

Name: Roman Kessler

Punkte:

Aufgabe 13. Verwenden Sie den R-Datensatz `faithful` (dieser Datensatz ist im base package enthalten). Das Merkmal `eruptions` enth?lt die L?nge 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 m?glichst aussagekr?ftig 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 m?ssen Sie die Optionen einstellen, damit das Histogramm als Dichtesch?tzer geplottet wird? Wie erh?lt man ein ?hnliches 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 Gau?- und Epanechnikov-Kernen.

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

L?sung

```
library(base)
df = data.frame(faithful)
```

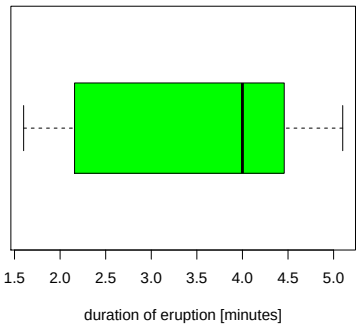
Zu (a):

Durch die Funktion “summary(dataFrame)” erhalten wir die 5 gesuchten Punkte und den arithmetischen Mittelwert:

```
summary(df)
#>    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):

```
boxplot(df$eruptions, col = "green", xlab = "duration of eruption [minutes]", horizontal = T
```



Wir sehen, dass die Dauer der Eruptionen zwischen etwa 1,5 und 5 Minuten ist, der Median jedoch bei etwa 4 Minuten liegt, was auf eine schiefe Verteilung der Daten schließen lässt.

Zu (c):

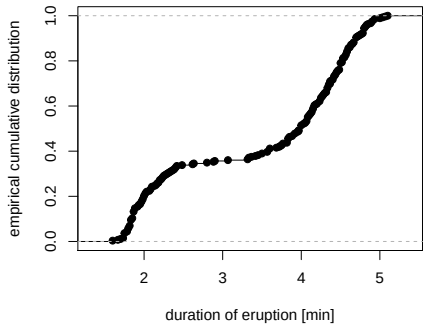
```
stem(df$eruptions)
#>
#>   The decimal point is 1 digit(s) to the left of the |
#>
#>   16 | 070355555588
#>   18 | 000022233333335577777777788882233577 7888
#>   20 | 000022233788000357778
#>   22 | 00023355780235 78
#>   24 | 00228
#>   26 | 23
#>   28 | 080
#>   30 | 7
#>   32 | 2337
#>   34 | 250077
#>   36 | 00008235777
#>   38 | 23333355822255777
```

```
#>    40 | 0000003357788888002233555577778
#>    42 | 033355557788002333335555777778
#>    44 | 022223355577800000000233333577778888
#>    46 | 00002333577000000023578
#>    48 | 00000022335800333
#>    50 | 0370
```

Der Stem-and-Leaf Plot gibt schon eine bessere Verteilung der Daten wieder. Wir sehen eine gewisse Dichotomie: Es gibt einige kurze Ausbrüche (um den Bereich von 2 Minuten) und viele längere Ausbrüche (um den Bereich von 3,5 bis 5 Minuten).

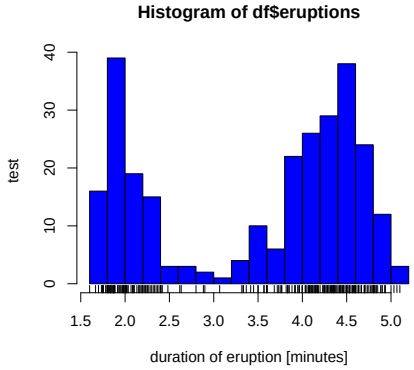
Zu (d):

```
plot(ecdf(df$eruptions), xlab = "duration of eruption [min]", ylab = "empirical cumulative d
```



Zu (e):

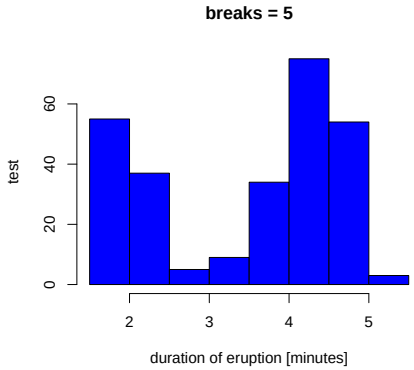
```
{
hist(df$eruptions, breaks = 20, col = "blue", xlab = "duration of eruption [minutes]", ylab
rug(df$eruptions)
}
```



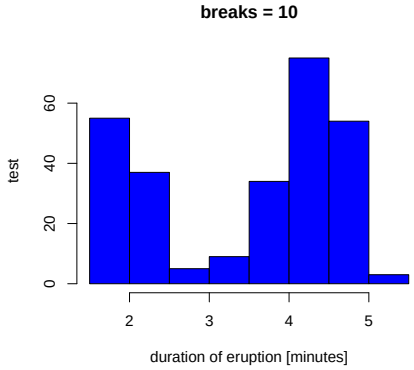
Das Histogramm (mit dem Rugplot) bestätigt noch einmal die Vermutung über die dichotome Verteilung der Daten.

Im folgenden experimentieren wir mit der Breite der "Bins".

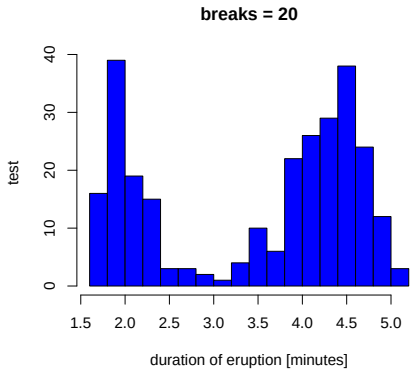
```
hist(df$eruptions, breaks = 5, col = "blue", xlab = "duration of eruption [minutes]", ylab =
```



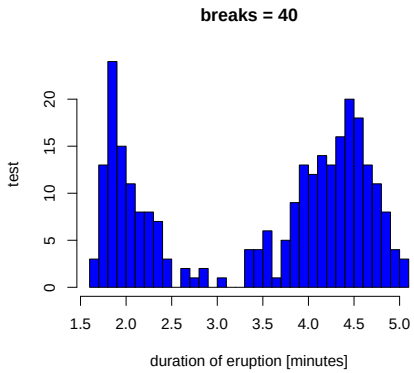

```
hist(df$eruptions, breaks = 10, col = "blue", xlab = "duration of eruption [minutes]", ylab
```



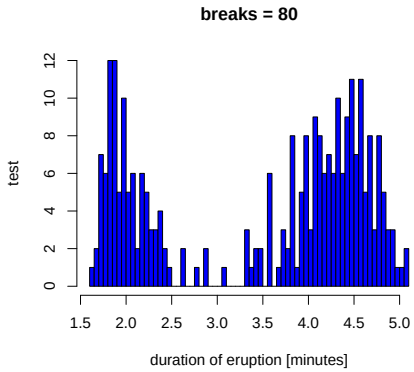

```
hist(df$eruptions, breaks = 20, col = "blue", xlab = "duration of eruption [minutes]", ylab
```



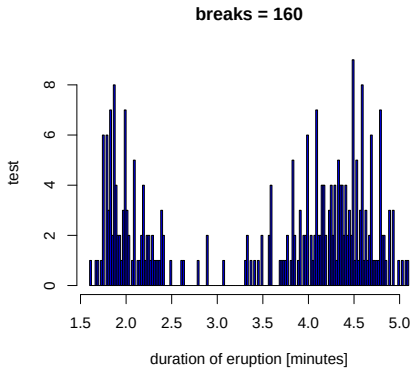

```
hist(df$eruptions, breaks = 40, col = "blue", xlab = "duration of eruption [minutes]", ylab
```



```
hist(df$eruptions, breaks = 80, col = "blue", xlab = "duration of eruption [minutes]", ylab
```




```
hist(df$eruptions, breaks = 160, col = "blue", xlab = "duration of eruption [minutes]", ylab
```



Wir sehen

1.: das Argument "breaks" (wenn wir ein Integer übergeben um die Anzahl der Zellen zu definieren), erst mal nur ein Vorschlag für die Histogram Funktion darstellt. Bei einem Argument (z.B. breaks = 5), wird dieser Vorschlag ignoriert, da das Programm (laut Dokumentation: "pretty") einen Wert festlegt, mit welchem ich als Betrachter die Verteilung der Daten besser dargestellt bekomme. Meine Eingabe (breaks = 5) scheint wohl eine ungünstiges Ergebnis zu liefern, welches mir eine andere Verteilung implizieren würde.

2.: Wähle ich die Bin-Anzahl zu hoch, liefert mir das keine "brauchbaren" Informationen über die generelle Verteilung der Daten. Eine Bin Größe, die der Gesamtzahl der Datenpunkte nahe kommt, ist somit weniger sinnvoll.

Zu (f):

```
library(aplpack)
slider.hist(df$eruptions, breaks = 5, col = "blue", xlab = "duration of eruption [minutes]",
#> Error in tkconfigure(sc, variable = slider1): konnte Funktion "tkconfigure" nicht finden
#?hist()
```

Zu (g):

Zu (h):

Anmerkungen/Korrektur
