Erste Schritte mit RStudio

Öffnen Sie RStudio

~/P2 - RStudio

File Edit Code View Plots Session Build Debug Profile Tools Help

Go to file/function | Addins | P2

QM3.R | P2_boxplots.R | source.R | R_useful.R | source2_.R | si

Source on Save | Run | Source

```
1  # so Schreiben Sie Kommentare
2
3
4  View(sib)
5  str(sib)
6  summary(sib)
7
8  ## Häufigkeitstabelle ##
9  tally(sib$number)  # absolute Häufigkeiten
10 freq<-tally(sib$number, format ="proportion") # relative Häufigkeit
11 freq
12 cumsum(freq)
13
14
```

82:16 (Top Level) R Script

Environment | History | Connections

Import Dataset | List

Global Environment

Data

Autos	5 obs. of 2 variables
chisq	List of 9
earth	6 obs. of 2 variables
grad	50 obs. of 1 variable
kg	9 obs. of 1 variable
NW	2 obs. of 2 variables
Schule	10 obs. of 2 variables

Files | Plots | Packages | Help | Viewer

Zoom | Export

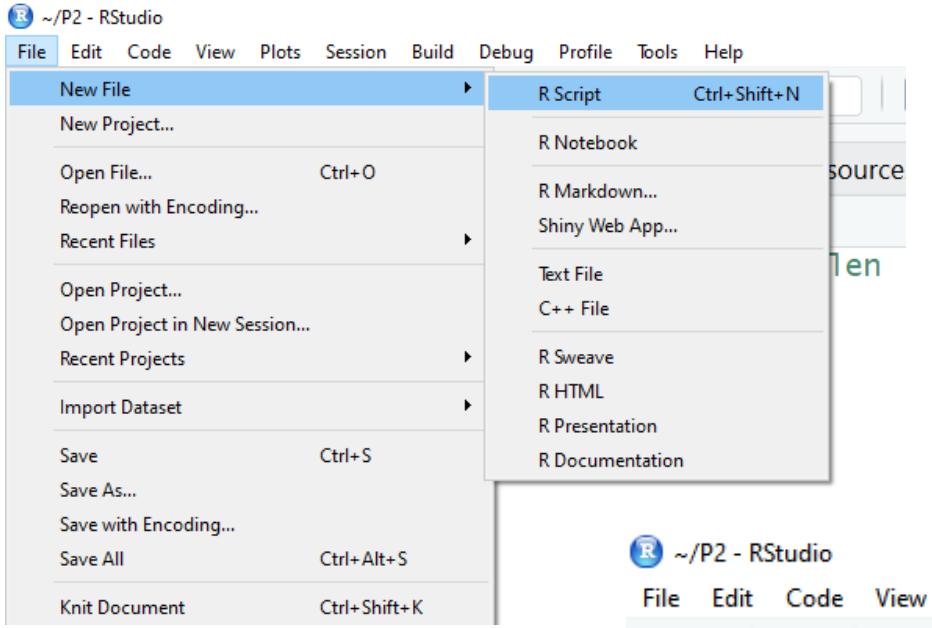
Console | Terminal

~/P2/

```
Type 'contributors()' for more information and
'citation()' on how to cite R or R packages in publications.

Type 'demo()' for some demos, 'help()' for on-line help, or
'help.start()' for an HTML browser interface to help.
Type 'q()' to quit R.

[Workspace loaded from ~/P2/.RData]

> library(mosaic)
```

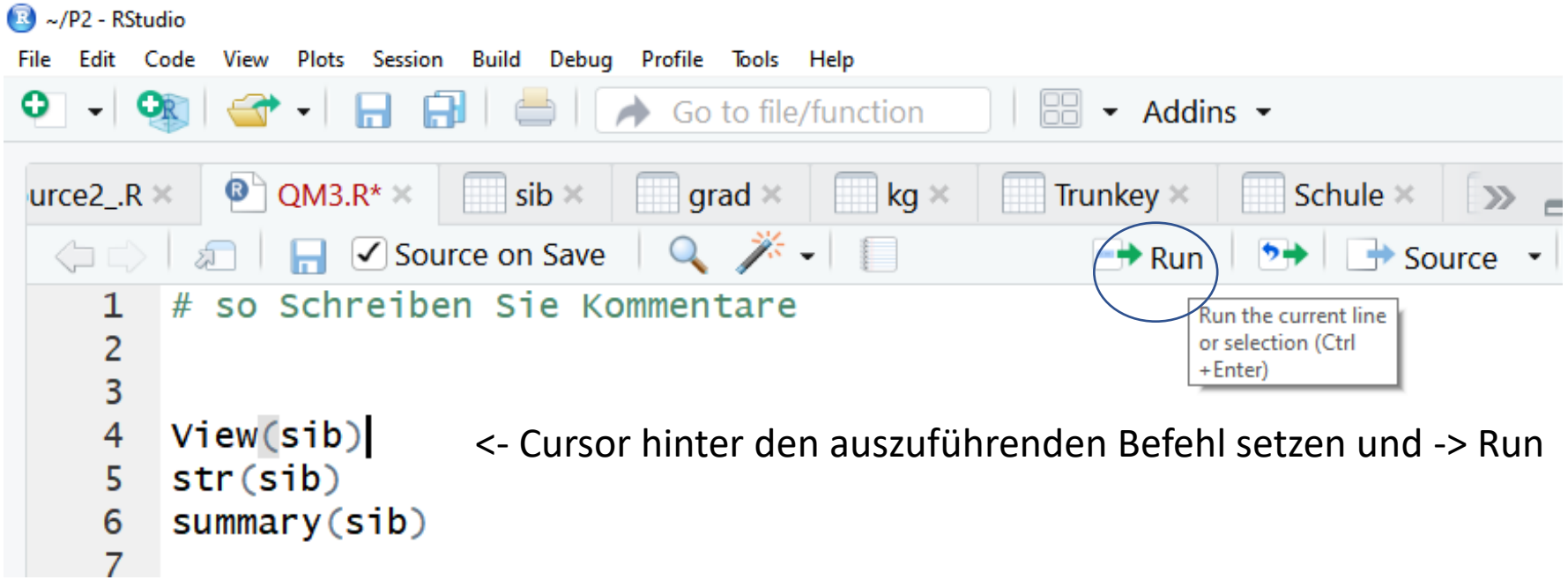
Importieren Sie mosaic

Ein R- Script anlegen und Befehle ausführen



Öffnen Sie ein leeres R Script (Endung ist .R) um hier alle Befehle fest zu halten.

File -> New File -> R Script
Save as (z.B.) „QM3“



<- Cursor hinter den auszuführenden Befehl setzen und -> Run

1.3 Häufigkeitstabelle: Beispiel

Absolute + relative Häufigkeiten

Es wurden 64 Personen nach der Anzahl ihrer Geschwister befragt. Die Urliste sei gegeben durch:

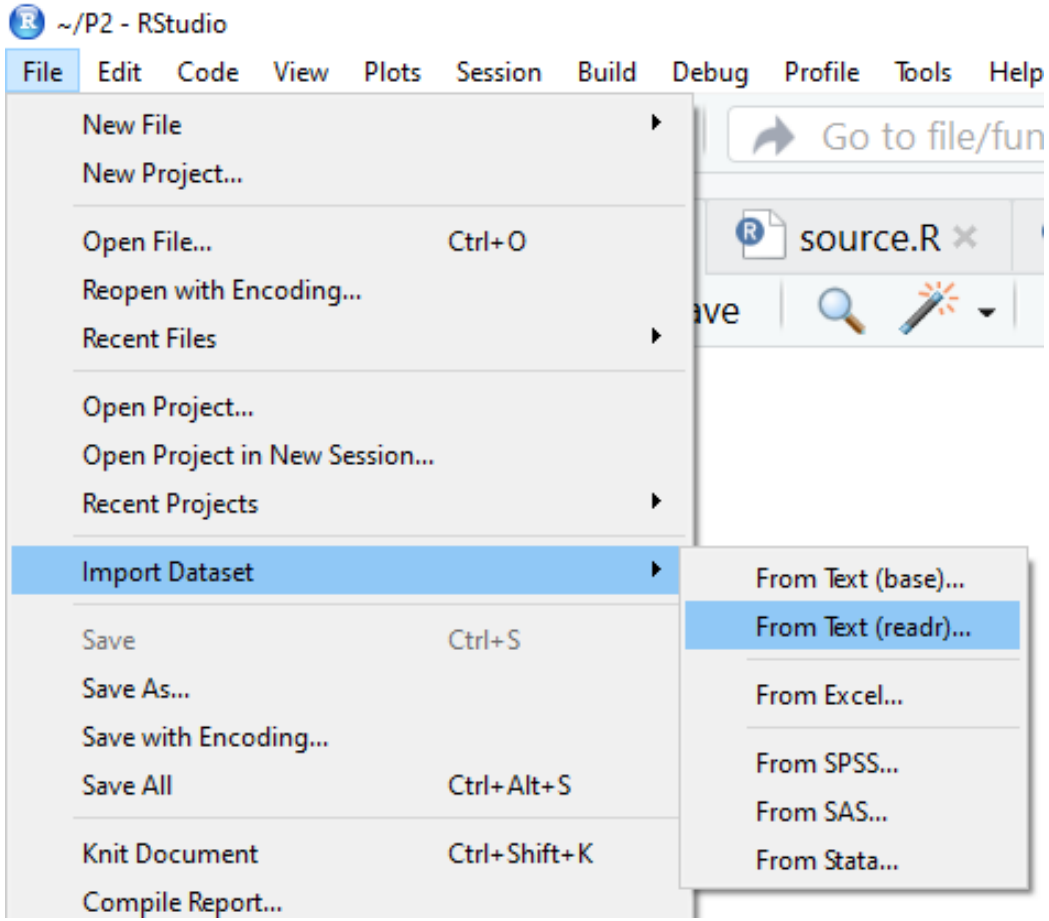
1 2 3 8 1 4 2 4 1 3 2 2 4 2 2 5 2 2 1 0 3 4 1 1 3 4 3 2 1 2 2 4
6 2 2 3 0 2 4 5 3 7 1 2 2 5 4 1 1 3 3 2 2 2 1 3 3 1 0 1 1 1 5 2

(a) Erstellen Sie die Häufigkeitstabelle. ← Wie kann ich die Daten übersichtlicher darstellen?

Nummer	Realisations-möglichkeit	abs. Häufigkeit	rel. Häufigkeit		kum. Häufigkeit	
1	0	3	$\frac{3}{64}$	0,0469	$\frac{3}{64}$	0,0469
2	1	15	$\frac{15}{64}$	0,2344	$\frac{18}{64}$	0,2813
3	2	20	$\frac{20}{64}$	0,3125	$\frac{38}{64}$	0,5938
4	3	11	$\frac{11}{64}$	0,1719	$\frac{49}{64}$	0,7656
5	4	8	$\frac{8}{64}$	0,1250	$\frac{57}{64}$	0,8906
6	5	4	$\frac{4}{64}$	0,0625	$\frac{61}{64}$	0,9531
7	6	1	$\frac{1}{64}$	0,0156	$\frac{62}{64}$	0,9688
8	7	1	$\frac{1}{64}$	0,0156	$\frac{63}{64}$	0,9844
9	8	1	$\frac{1}{64}$	0,0156	1	1,0000

$$rel.H. = \frac{abs.H.}{Anzahl}$$

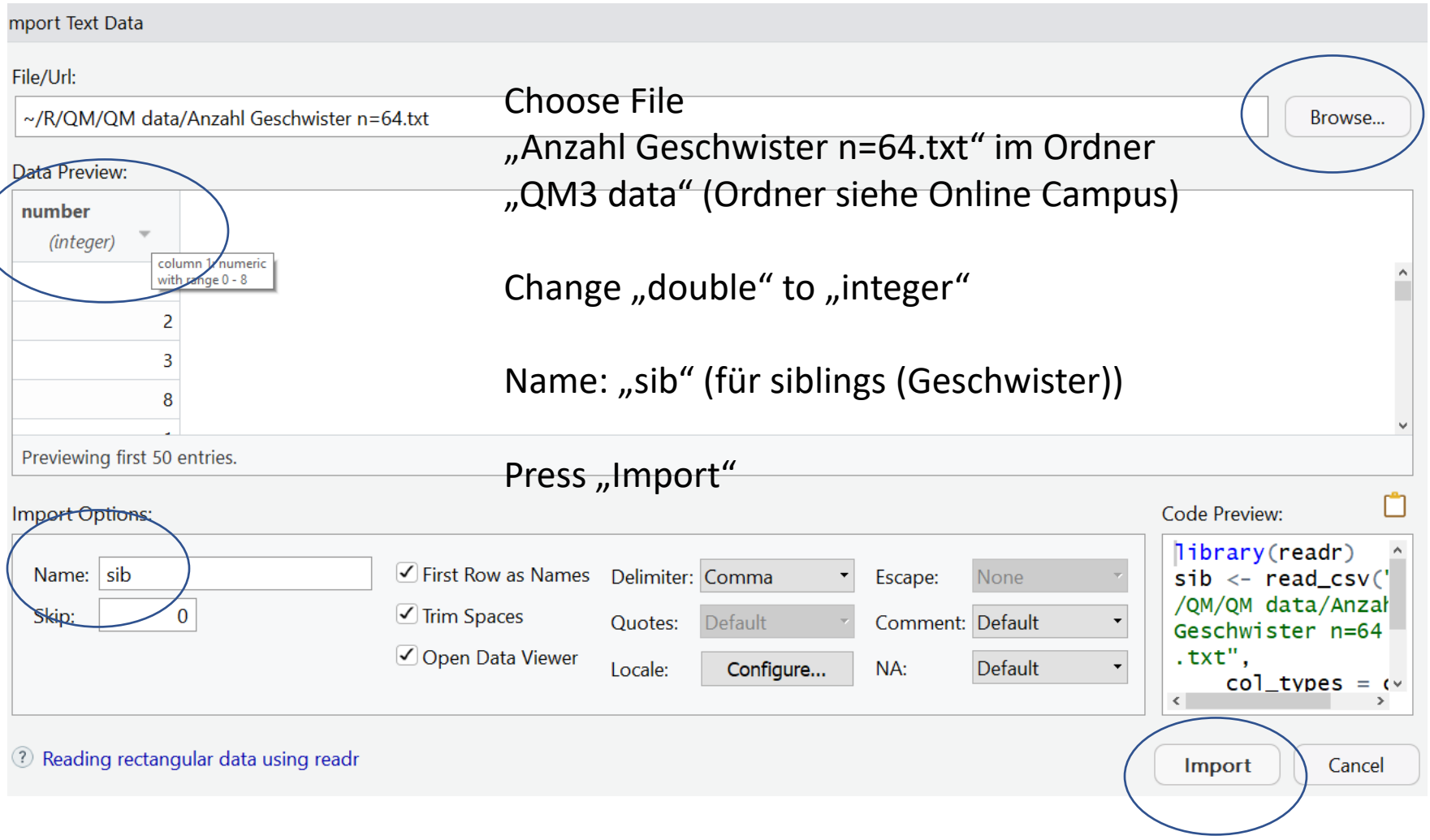
Importieren von .txt- Dateien



File -> Import Dataset -> From Text (readr)
Es öffnet sich das Import Tool (nächste Folie)

Import Tool

Ordner „QM3 data“siehe Online Campus

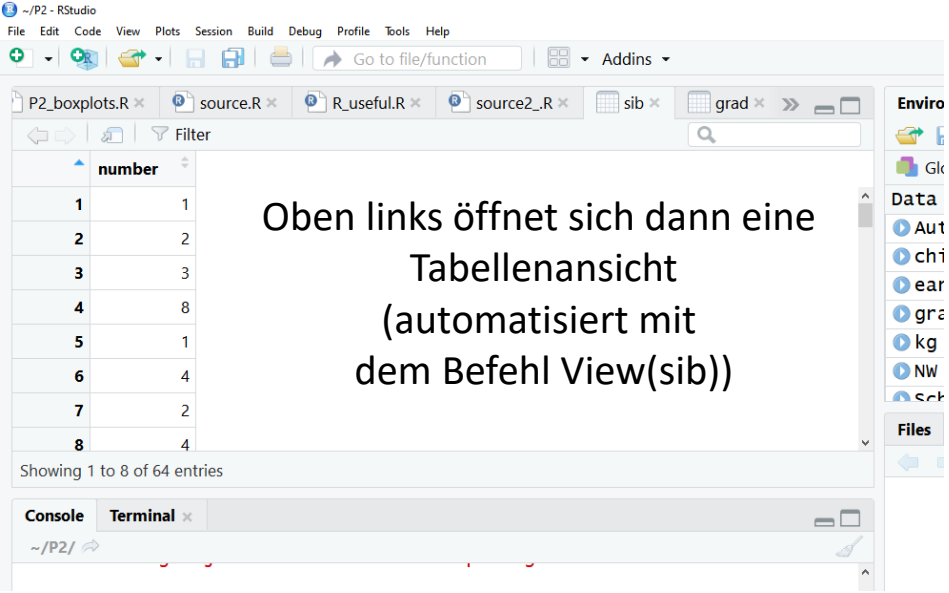


Choose File
„Anzahl Geschwister n=64.txt“ im Ordner „QM3 data“ (Ordner siehe Online Campus)

Change „double“ to „integer“

Name: „sib“ (für siblings (Geschwister))

Press „Import“



Oben links öffnet sich dann eine Tabellenansicht (automatisiert mit dem Befehl View(sib))

```
6  tally(sib$number)  # absolute Häufigkeiten
7  freq<-tally(sib$number, format ="proportion") # relative Häufigkeit
8  freq
9  cumsum(freq)
10
```

5:1 (Top Level) R Script

Console Terminal

~/P2/

```
> tally(sib$number)  # absolute Häufigkeiten
X
 0  1  2  3  4  5  6  7  8
 3 15 20 11  8  4  1  1  1
> freq<-tally(sib$number, format ="proportion") # relative Häufigkeiten
> freq
X
       0        1        2        3        4        5        6
0.046875 0.234375 0.312500 0.171875 0.125000 0.062500 0.015625
       7        8
0.015625 0.015625
> cumsum(freq)
       0        1        2        3        4        5        6
0.046875 0.281250 0.593750 0.765625 0.890625 0.953125 0.968750
       7        8
0.984375 1.000000
>
```

Aufgabe 1

Es wurden 64 Personen nach der Anzahl ihrer Geschwister befragt. Die Urliste sei gegeben durch:

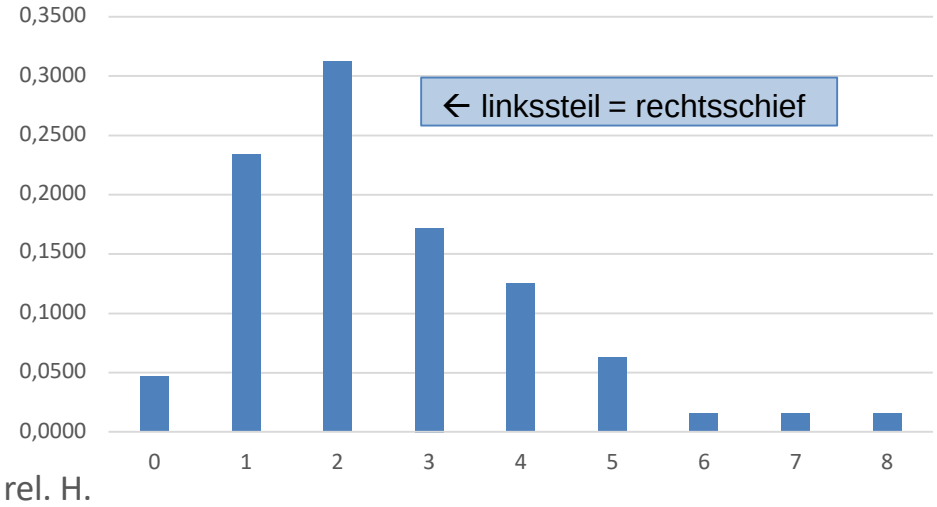
1 2 3 8 1 4 2 4 1 3 2 2 4 2 2 5 2 2 1 0 3 4 1 1 3 4 3 2 1 2 2 4
6 2 2 3 0 2 4 5 3 7 1 2 2 5 4 1 1 3 3 2 2 2 1 3 3 1 0 1 1 1 5 2

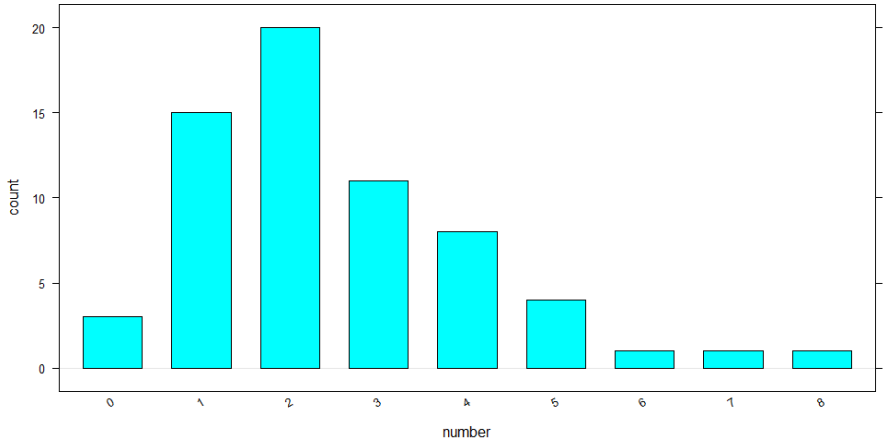
(a) Erstellen Sie die Häufigkeitstabelle.

Nummer	Realisations-möglichkeit	abs. Häufigkeit	rel. Häufigkeit		kum. Häufigkeit	
1	0	3	$\frac{3}{64}$	0,0469	$\frac{3}{64}$	0,0469
2	1	15	$\frac{15}{64}$	0,2344	$\frac{18}{64}$	0,2813
3	2	20	$\frac{20}{64}$	0,3125	$\frac{38}{64}$	0,5938
4	3	11	$\frac{11}{64}$	0,1719	$\frac{49}{64}$	0,7656
5	4	8	$\frac{8}{64}$	0,1250	$\frac{57}{64}$	0,8906
6	5	4	$\frac{4}{64}$	0,0625	$\frac{61}{64}$	0,9531
7	6	1	$\frac{1}{64}$	0,0156	$\frac{62}{64}$	0,9688
8	7	1	$\frac{1}{64}$	0,0156	$\frac{63}{64}$	0,9844
9	8	1	$\frac{1}{64}$	0,0156	1	1,0000

Säulendiagramm, Empirische Verteilungsfunktion und Histogramm

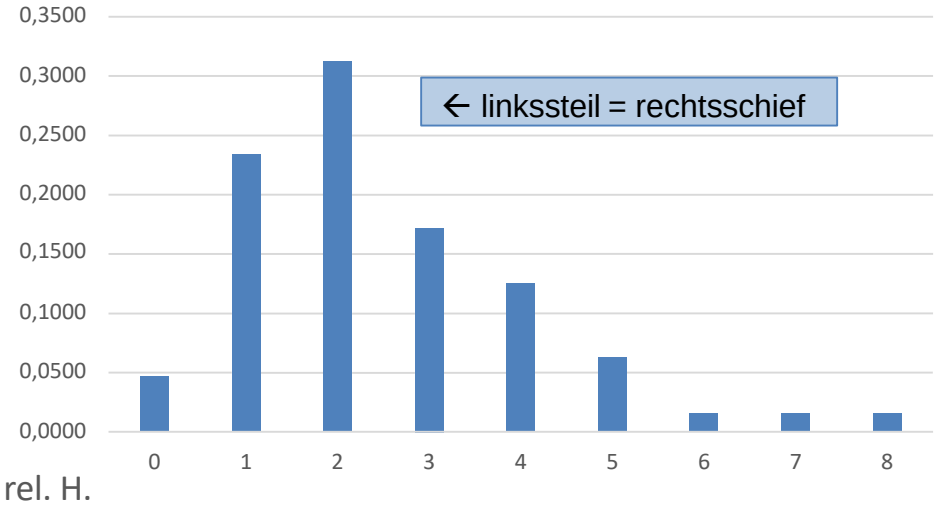
1.4 Stabdiagramm: siblings mit R

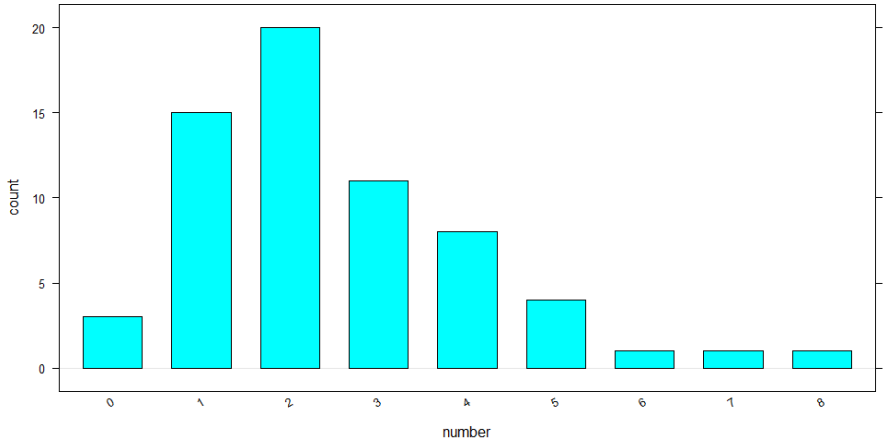





```
bargraph(~number, data=sib)
```

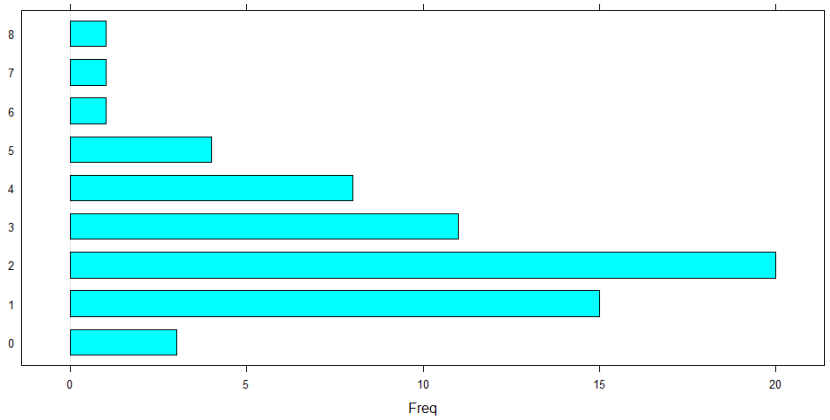
1.4 Stabdiagramm: siblings mit R



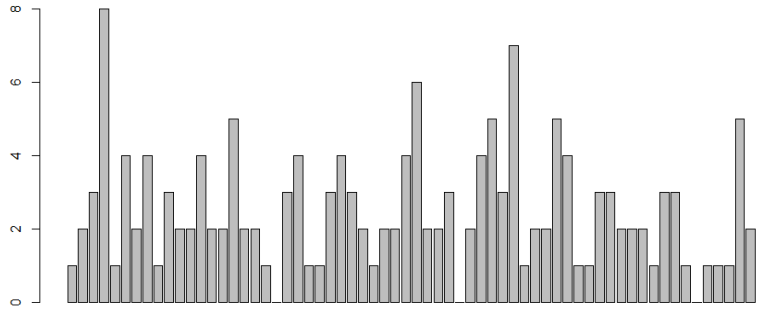



```
bargraph(~number, data=sib)
```

Andere:



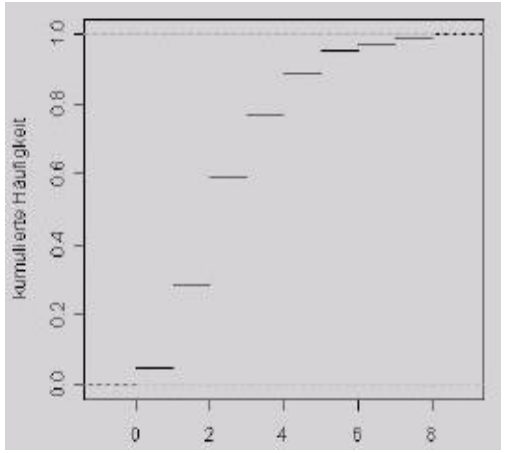

```
barchart(sib) # Balkendiagramm
```

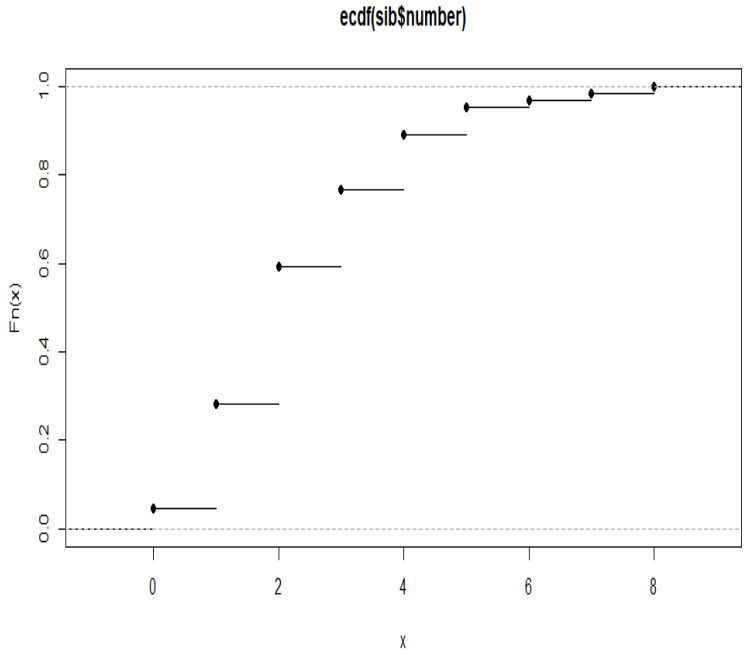


```
barplot(sib$number)
```

1.5 Empirische Verteilungsfunktion

Merkmals-Ausprägung x_i	*kumulative Häufigkeit* $\sum_{i=1}^{n} h_i$
0	0,0469
1	0,2813
2	0,5938
3	0,7656
4	0,8906
5	0,9531
6	0,9688
7	0,9844
8	1,0000






```
plot(ecdf(sib$number))
```

ecdf = empirical cumulative distribution function

(d) Wie groß ist der Anteil der Personen, die genau 3 Geschwister haben?

Der Anteil der Personen, die drei Geschwister haben, kann aus der Häufigkeitstabelle abgelesen werden:

$$h(X=3) = \frac{11}{64} = 0,172$$

17,2 % der Personen haben drei Geschwister.

Nummer	Realisations-möglichkeit	abs. Häufigkeit	rel. Häufigkeit	kum. Häufigkeit
1	0	3	$\frac{3}{64}$	$\frac{3}{64}$
2	1	15	$\frac{15}{64}$	$\frac{18}{64}$
3	2	20	$\frac{20}{64}$	$\frac{38}{64}$
4	3	11	$\frac{11}{64}$	$\frac{49}{64}$
5	4	8	$\frac{8}{64}$	$\frac{57}{64}$
6	5	4	$\frac{4}{64}$	$\frac{61}{64}$
		1	$\frac{1}{64}$	$\frac{62}{64}$
		1	$\frac{1}{64}$	$\frac{63}{64}$
		1	$\frac{1}{64}$	1

← Erkenntnisse gewinnen

Console Terminal

~/P2/

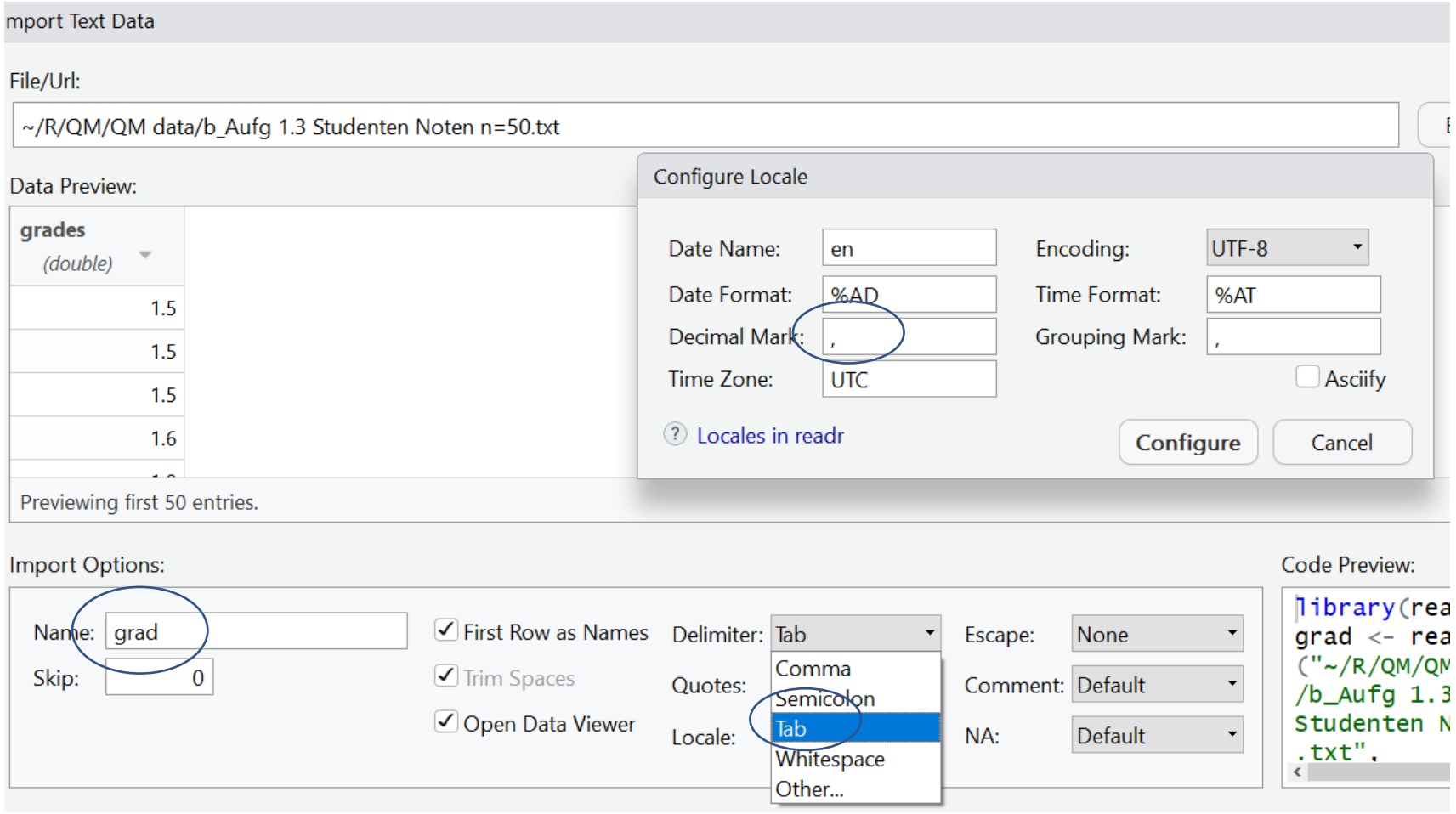
```
> freq
X
       0        1        2        3        4        5        6        7
0.046875 0.234375 0.312500 0.171875 0.125000 0.062500 0.015625 0.015625
       8
0.015625
> freq[3]
     2
0.3125
> freq[4]
       3
0.171875
>
```

Beispieldatensatz „Studenten Noten n=50“

Aus den Fragebögen wurden 50 Studenten ausgewählt und die Schulabschlussnoten notiert. Daraus ergibt sich folgende Urliste:

2.1	2.1	2.2	2.5	3.0	2.0	2.2	2.4	2.9	2.4
3.6	3.1	3.2	2.3	3.3	1.8	2.7	3.7	2.9	2.9
3.2	3.4	2.3	2.4	3.0	2.0	2.6	3.5	3.2	3.0
3.4	2.6	2.5	1.5	1.5	3.0	2.0	2.5	2.9	2.8
1.6	2.0	2.6	2.1	3.2	3.0	3.5	2.4	1.5	3.3

Import „Studenten Noten n=50“
Delimiter: Tab
Decimal Mark: Comma
Type: double
Name: grad



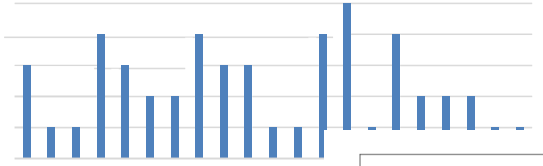
Beispiel

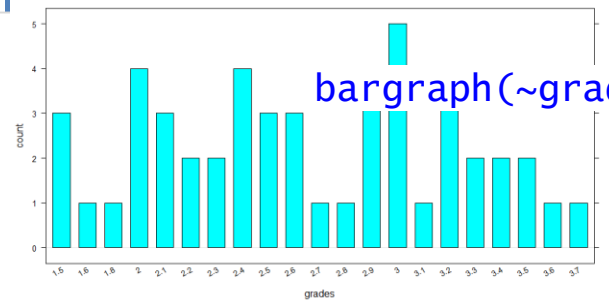
Häufigkeitstabelle

Note	abs.H.	rel.H.	kum.H.
1,5	3	0,060	0,060
1,6	1	0,020	0,080
1,8	1	0,020	0,100
2	4	0,080	0,180
2,1	3	0,060	0,240
2,2	2	0,040	0,280
2,3	2	0,040	0,320
2,4	4	0,080	0,400
2,5	3	0,060	0,460
2,6	3	0,060	0,520
2,7	1	0,020	0,540
2,8	1	0,020	0,560
2,9	4	0,080	0,640
3	5	0,100	0,740
3,1	1	0,020	0,760
3,2	4	0,080	0,840
3,3	2	0,040	0,880
3,4	2	0,040	0,920
3,5	2	0,040	0,960
3,6	1	0,020	0,980
3,7	1	0,020	1,000

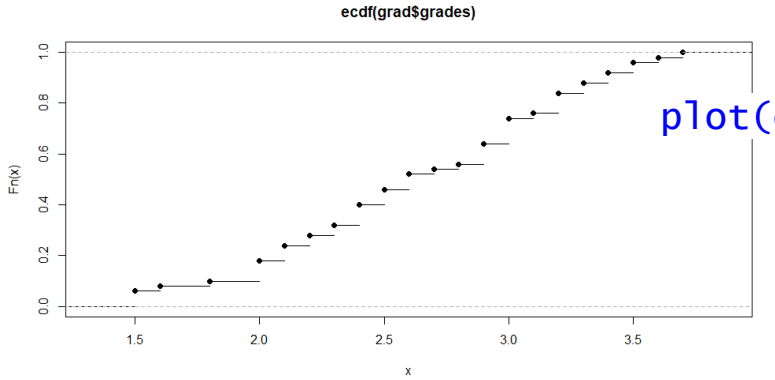
n=50

Stabdiagramm mit rel.H.



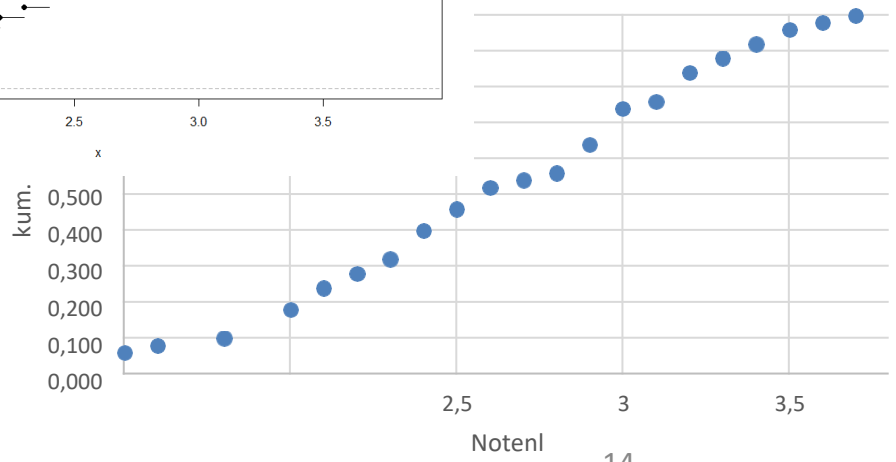



```
bargraph(~grades, data=grad)
```

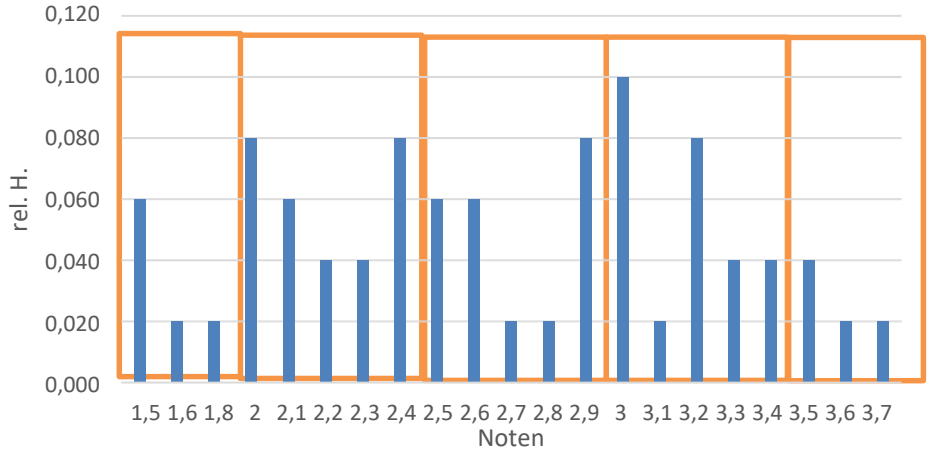



```
plot(ecdf(grad$grades))
```

empirische Vert.-fkt.



1.6 Klassenbildung und Histogramm

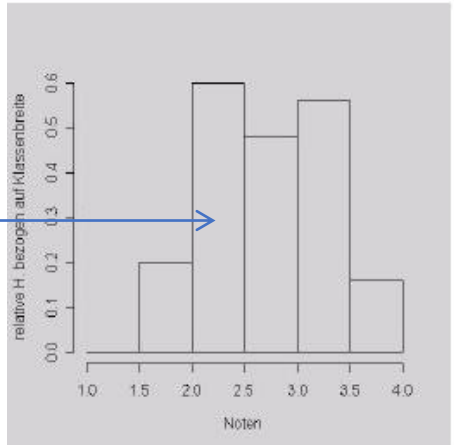


(a) Erstellen Sie die Häufigkeitstabelle.

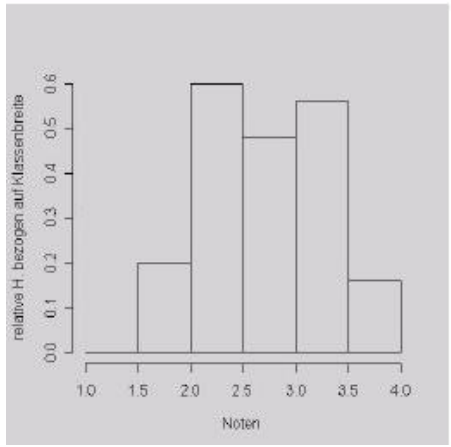
Klasse	Intervall	abs. Häufigkeit	rel. Häufigkeit
1	$[1; 1,5)$	0	0
2	$[1,5; 2)$	5	0,1
3	$[2; 2,5)$	15	0,3
4	$[2,5; 3)$	12	0,24
5	$[3; 3,5)$	14	0,28
6	$[3,5; 4)$	4	0,08

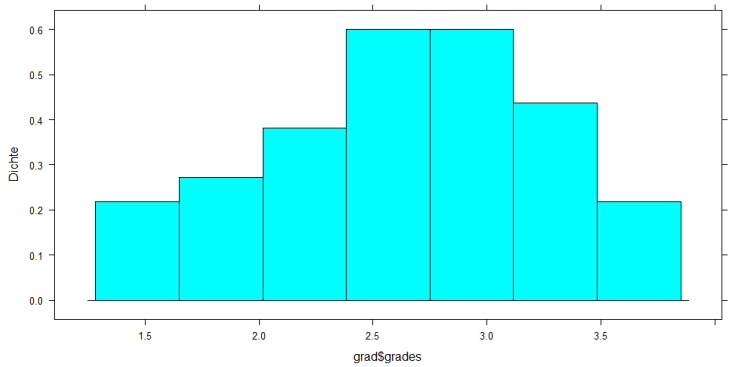
0,5 * 0,6 = 0,3

Halboffenes Intervall: die 3,5 ist mit drin, die 4 nicht mehr



1.6 Klassenbildung und Histogramm






```
histogram(grad$grades)
```

default Histogramm

Lagemaße (MW, Median und Modus),
Boxplots und Quantile

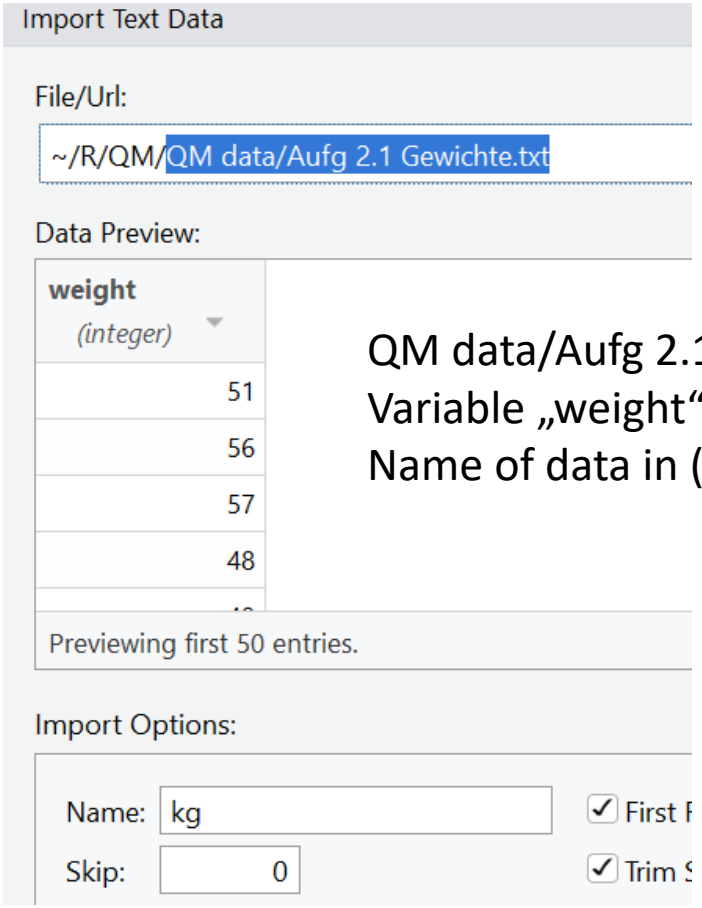
2.1 Mittelwert und Median

Aus einem Fragebogen wurde als interessierendes Merkmal das Gewicht ausgewählt. Hierzu wurde das Gewicht von 9 Personen aufgelistet:

51 56 57 48 49 61 46 50 59

(a) Bestimmen Sie den Mittelwert und den Median des Merkmals Gewicht.

$$\bar{x} = \frac{1}{9}(46 + 48 + 49 + ... + 59 + 61) = 53$$



QM data/Aufg 2.1 Gewichte.txt
Variable „weight“ as integer einlesen
Name of data in (z.B.) „kg“ umbenennen

2.2 Boxplot

(c) Erstellen und interpretieren Sie den Boxplot der Daten.

5-Zahlen-Zusammenfassung:

$$x_{(1)} = 46 \quad x_{0.25} = 49 \quad x_{0.5} = 51 \quad x_{0.75} = 57 \quad x_{(n)} = 61$$

Ungerade Anzahl an n:

46(Min) Q_{25} (49) 51 (Median) Q_{75} (57) 61 (Max)

i	1	2	3	4	5	6	7	8	9
x_i	46	48	49	50	51	56	57	59	61

```
> mean(kg$weight) # arithmetischer Mittelwert
[1] 53
> median(kg$weight) # Median
[1] 51
> min(kg$weight)
[1] 46
> max(kg$weight)
[1] 61
> quantile(kg$weight)
  0%  25%  50%  75% 100%
  46   49   51   57   61
> iqr(kg$weight)
[1] 8
```

ODER

```
> favstats(kg$weight)
 min Q1 median Q3 max mean       sd n missing
  46 49     51 57  61   53 5.338539 9       0
```

Gerade Anzahl an n

i	1	2	3	4	5	6	7	8	9	10
x_i	46	48	49	50	51	56	57	59	61	61

Interquartilsabstand $iqr = Q_{75} - Q_{25}$

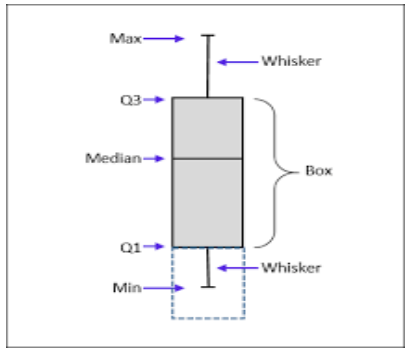
$$\bar{x} = \frac{51 + 56}{2} = 53{,}5 \text{ (Median)}$$

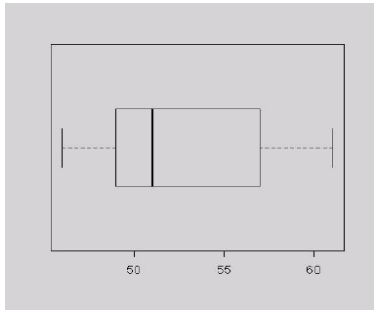
2.2 Boxplot

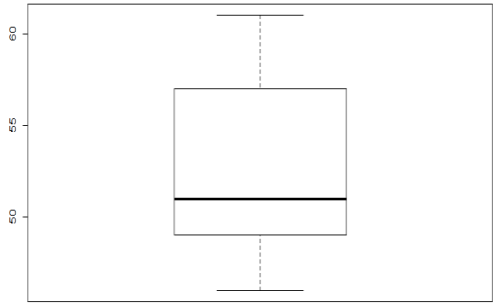
(c) Erstellen und interpretieren Sie den Boxplot der Daten.

5-Zahlen-Zusammenfassung:

$$x_{(1)} = 46 \quad x_{0.25} = 49 \quad x_{0.5} = 51 \quad x_{0.75} = 57 \quad x_{(n)} = 61$$

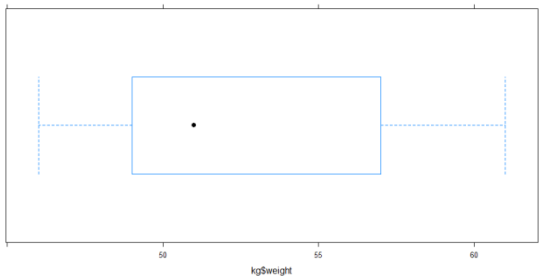




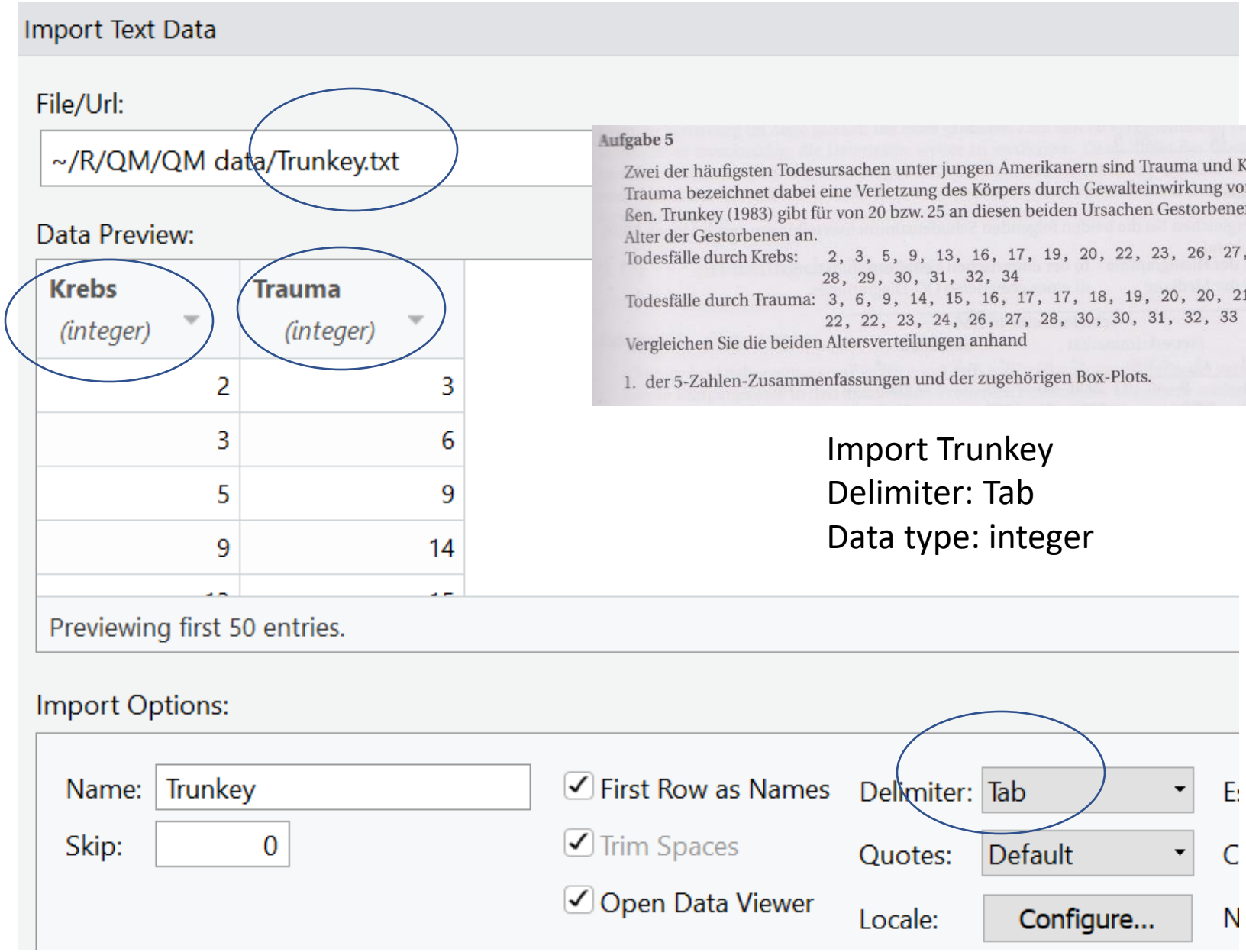



```
boxplot(kg$weight)
```

Andere:




```
bwplot(kg$weight)
#horizontaler boxplot
```



Aufgabe 5

Zwei der häufigsten Todesursachen unter jungen Amerikanern sind Trauma und Krebs. Trauma bezeichnet dabei eine Verletzung des Körpers durch Gewalteinwirkung von außen. Trunkey (1983) gibt für von 20 bzw. 25 an diesen beiden Ursachen Gestorbenen das Alter der Gestorbenen an.

Todesfälle durch Krebs: 2, 3, 5, 9, 13, 16, 17, 19, 20, 22, 23, 26, 27, 27, 28, 29, 30, 31, 32, 34

Todesfälle durch Trauma: 3, 6, 9, 14, 15, 16, 17, 17, 18, 19, 20, 20, 21, 22, 22, 23, 24, 26, 27, 28, 30, 30, 31, 32, 33

Vergleichen Sie die beiden Altersverteilungen anhand

1. der 5-Zahlen-Zusammenfassungen und der zugehörigen Box-Plots.

Import Trunkey
Delimiter: Tab
Data type: integer

Boxplot Beispiel

Aufgabe 5

Zwei der häufigsten Todesursachen unter jungen Amerikanern sind Trauma und Krebs. Trauma bezeichnet dabei eine Verletzung des Körpers durch Gewalteinwirkung von außen. Trunkey (1983) gibt für von 20 bzw. 25 an diesen beiden Ursachen Gestorbenen das Alter der Gestorbenen an.

Todesfälle durch Krebs: 2, 3, 5, 9, 13, 16, 17, 19, 20, 22, 23, 26, 27, 27, 28, 29, 30, 31, 32, 34

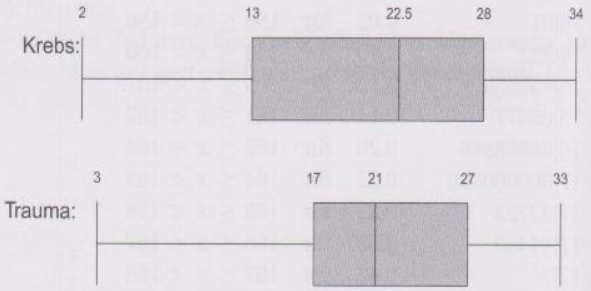
Todesfälle durch Trauma: 3, 6, 9, 14, 15, 16, 17, 17, 18, 19, 20, 20, 21, 22, 22, 23, 24, 26, 27, 28, 30, 30, 31, 32, 33

Vergleichen Sie die beiden Altersverteilungen anhand

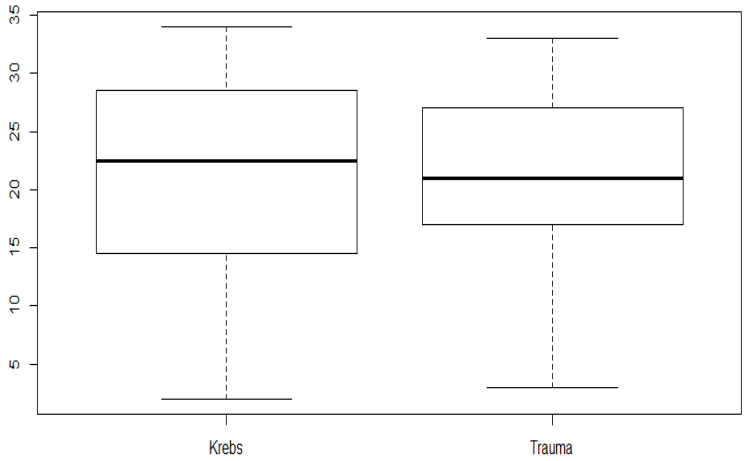
1. der 5-Zahlen-Zusammenfassungen und der zugehörigen Box-Plots.

1. Die 5-Zahlen Zusammenfassungen sind:

	Krebs ($n=20$)	Trauma ($n=25$)
$x_{(1)}$	2	3
$x_{0.25}$	13	17
$\tilde{x}$	22.5	21
$x_{0.75}$	28	27
$x_{(n)}$	34	33




```
> favstats(Trunkey$Krebs)
 min    Q1 median    Q3 max  mean       sd  n missing
   2 15.25   22.5 28.25  34 20.65 9.943392 20       5
> favstats(Trunkey$Trauma)
 min Q1 median Q3 max  mean       sd  n missing
   3 17     21 27  33 20.92 7.910541 25       0
```




```
boxplot(Trunkey)
```

Quelle: Schlittgen, „Einführung in die Statistik“ S. 460