

Statistik für Informatiker, SS 2019

2. Ideen aus der Statistik

2.1 Deskriptive Statistik

Matthias Birkner

<http://www.staff.uni-mainz.de/birkner/StatInfo19/>

24.6.2019



JOHANNES GUTENBERG
UNIVERSITÄT MAINZ

Viele Menschen stehen „Statistik“ kritisch gegenüber:

*It is easy to lie with statistics.
It is hard to tell the truth without it.*

Andrejs Dunkels (1939–1998)

Worum geht es in der Statistik?

Die Welt ist voller Variabilität.

Wie geht man mit variablen Daten um?

Idee der Statistik:

Variabilität (Erscheinung der Natur) durch Zufall
(mathematische Abstraktion) modellieren

Die Daten werden als Realisierungen von Zufallsvariablen
aufgefasst, die in einem stochastischen Modell spezifiziert
werden.

Man versucht dann, anhand der Daten Rückschlüsse auf
Parameter des Modells zu ziehen, und so systematische
Effekte von Zufälligem zu trennen.

Deskriptive (d.h. beschreibende) Statistik

Wie geht man mit variablen Daten um?

„0. Antwort“: Man verschafft sich einen ersten Eindruck mittels graphischer Darstellungen und statistischer Kenngrößen

Ein Beispiel

Bei einer biologischen Expedition wurden in der Nordsee Springkrebse (*Galathea intermedia*) gefangen und untersucht.

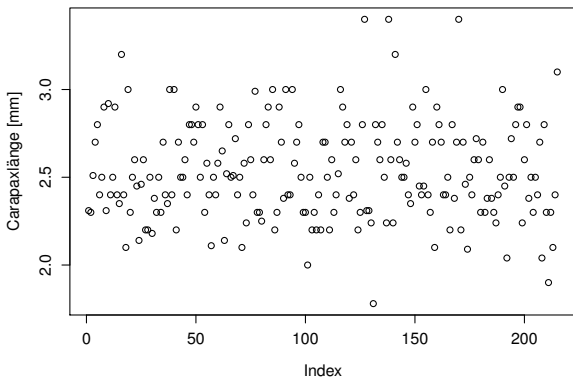


Die Daten: Helgoländer Tiefe Rinne, Fang vom 6.9.

Carapaxlänge (mm):

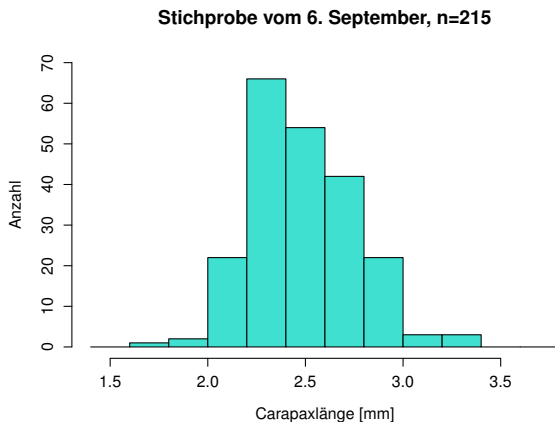
Nichteiertragende Weibchen ($n = 215$)

2,9	3,0	2,9	2,5	2,7	2,9	2,9	3,0
3,0	2,9	3,4	2,8	2,9	2,8	2,8	2,4
2,8	2,5	2,7	3,0	2,9	3,2	3,1	3,0
2,7	2,5	3,0	2,8	2,8	2,8	2,7	3,0
2,6	3,0	2,9	2,8	2,9	2,9	2,3	2,7
2,6	2,7	2,5

Stichprobe vom 6. September, n=215

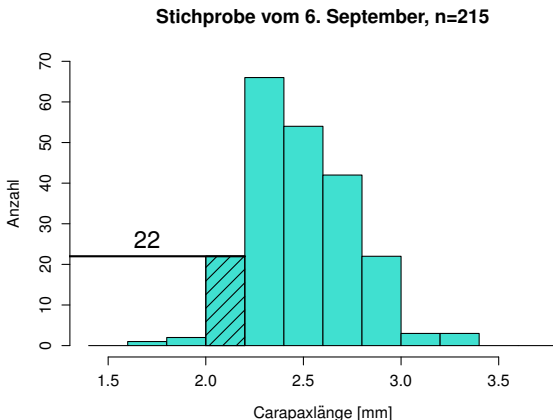
Eine Möglichkeit der graphischen Darstellung: das Histogramm

Histogramm der Carapaxlängen in der Stichprobe



Wieviele hatten Carapaxlänge zwischen 2,0 und 2,2 mm ?

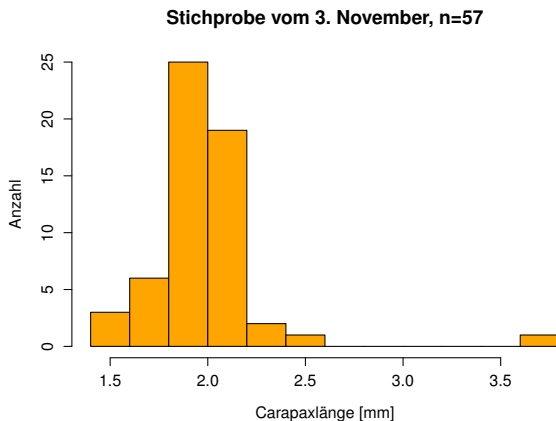
Histogramm der Carapaxlängen in der Stichprobe



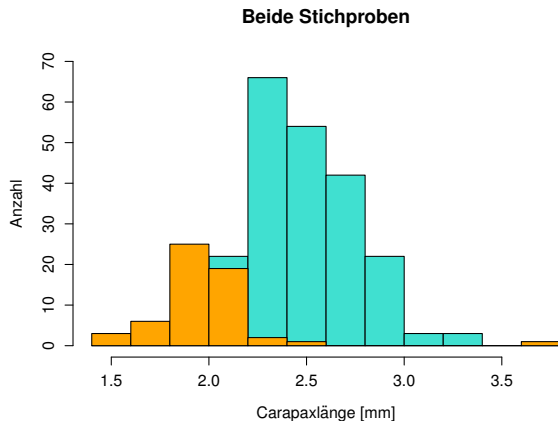
Wieviele hatten Carapaxlänge zwischen 2,0 und 2,2 mm ?
22

Analoge Daten zwei Monate später

(Stichprobe vom 3.11. der Größe $n = 57$)



Vergleich der beiden Verteilungen



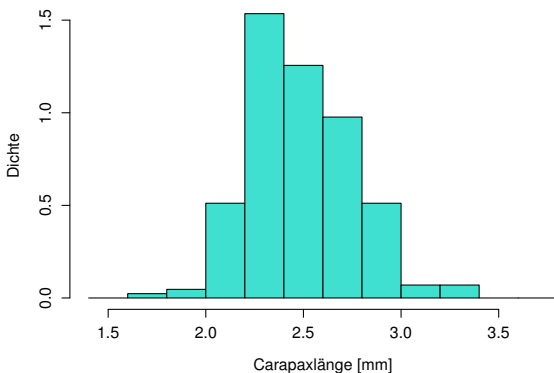
Problem: ungleiche Stichprobenumfänge:

6.Sept: $n = 215$

3.Nov : $n = 57$

Idee: stauche vertikale Achse so, dass Gesamtfläche = 1.

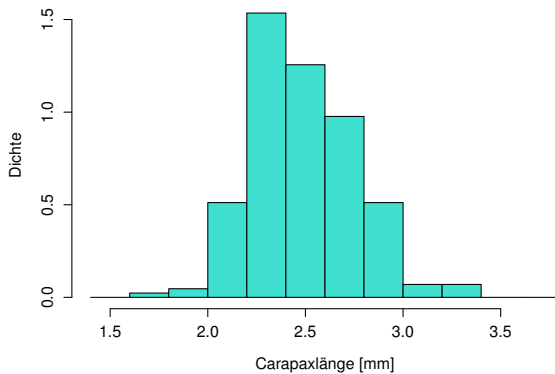
Stichprobe vom 6. September, n=215



Die Gesamtfläche der Balken ist nun $= 1$.

Die neue vertikale Koordinate ist jetzt eine **Dichte** (engl. **density**).

Stichprobe vom 6. September, n=215

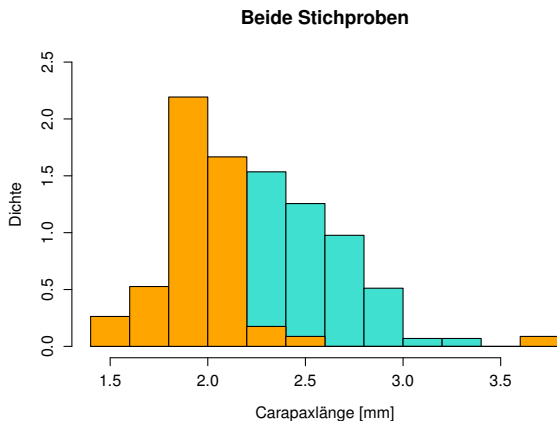


Stichprobe vom 6. September, n=215

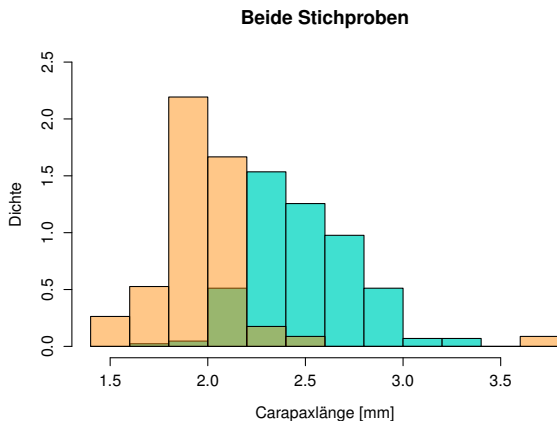


Die beiden Histogramme sind jetzt
vergleichbar
(sie haben dieselbe Gesamtfläche).

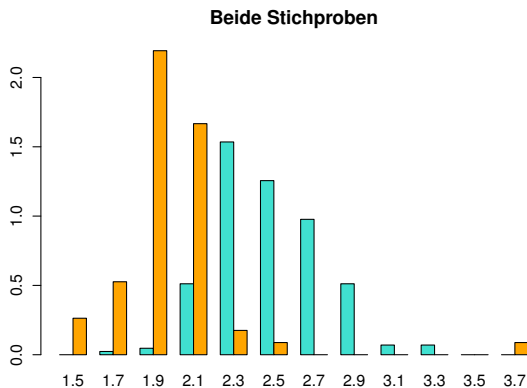
Versuche, die Histogramme zusammen zu zeigen:



Versuche, die Histogramme zusammen zu zeigen:

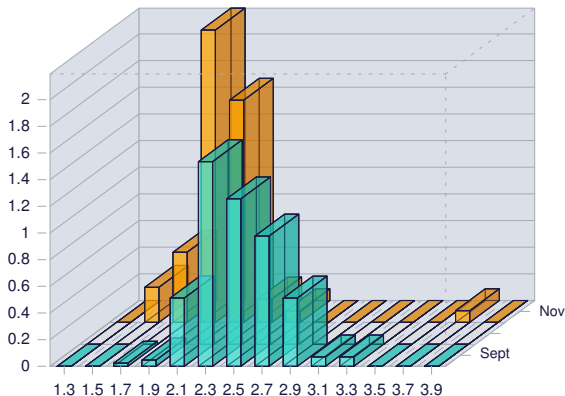


Versuche, die Histogramme zusammen zu zeigen:



Versuche, die Histogramme zusammen zu zeigen:

Beide Stichproben



Vorschlag

Total abgefahrene 3D-Plots können in der Werbung nützlich sein,

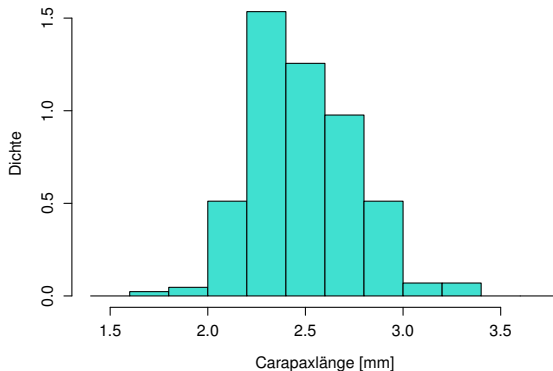
für die Wissenschaft sind einfache und klare
2D-Darstellungen meistens angemessener.

Problem

Histogramme kann man nicht ohne
weiteres
in demselben Graphen
darstellen,
weil sie einander
überdecken würden.

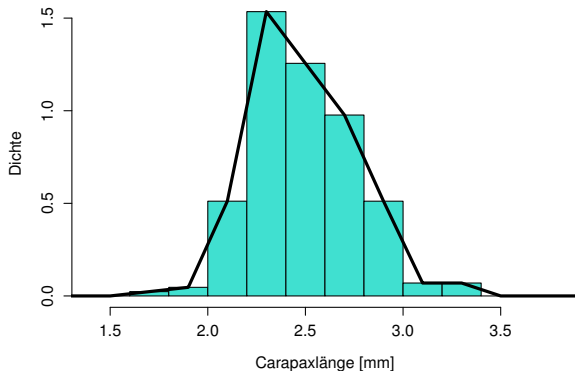
Einfache und klare Lösung: Dichtepolygone

Stichprobe vom 6. September, n=215

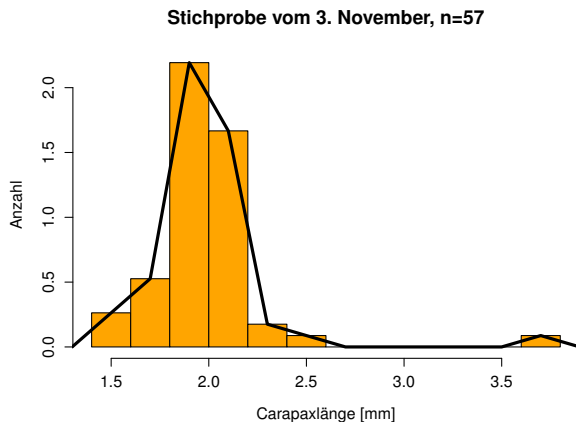


Einfache und klare Lösung: Dichtepolygone

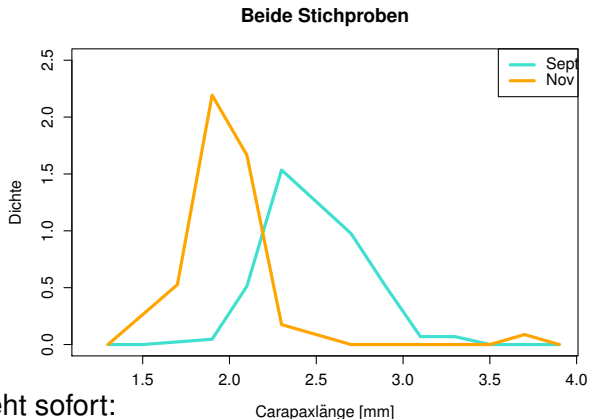
Stichprobe vom 6. September, n=215



Einfache und klare Lösung: Dichtepolygone



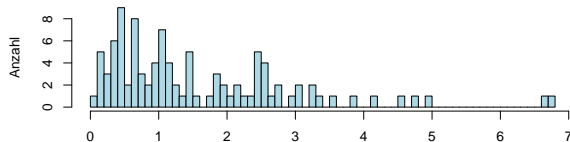
Zwei (oder mehr) Dichtepolygone in einem Plot



Man sieht sofort:

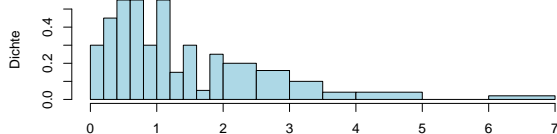
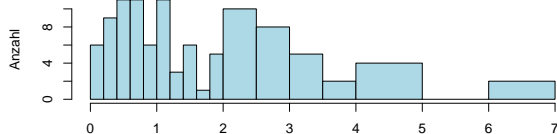
Die Verteilung in der Stichprobe vom November ist gegenüber der vom September nach links verschoben (und sie ist auch stärker um den häufigsten Wert konzentriert).

Anzahl vs. Dichte



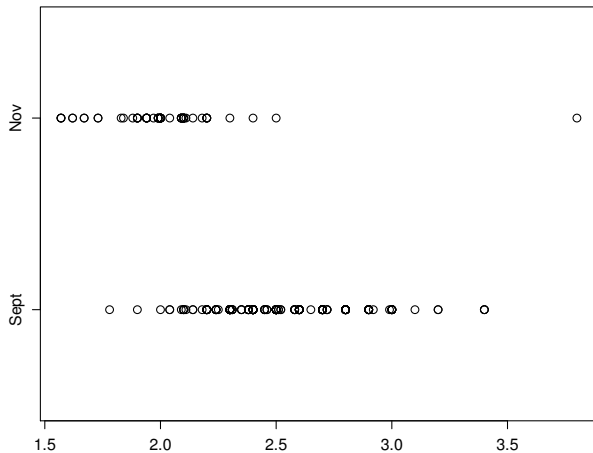
Also:

Bei Histogrammen mit ungleichmäßiger Unterteilung immer Dichten verwenden!



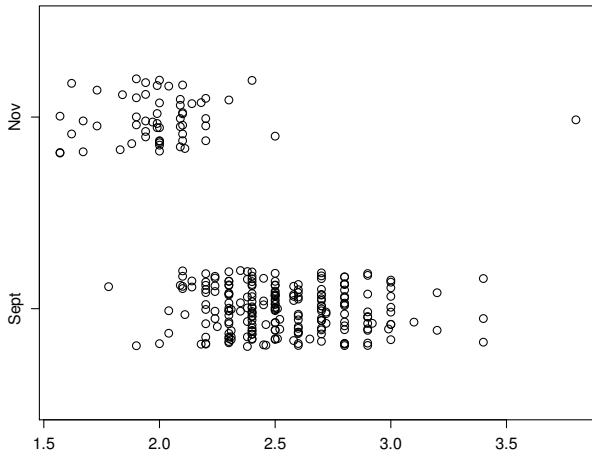
Stripchart, einfach

Carapaxlängen in den beiden Stichproben



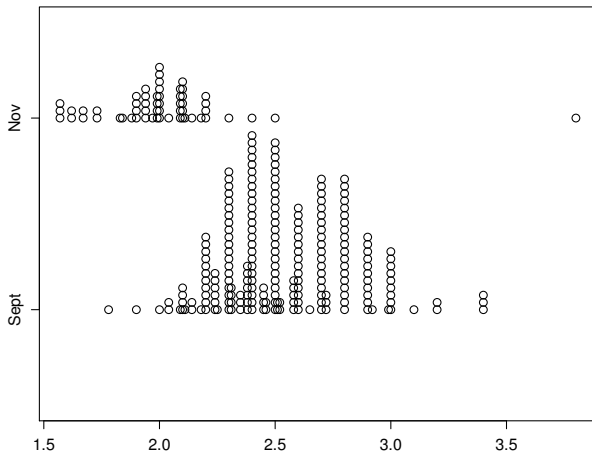
Stripchart, mit "jitter"

Carapaxlängen in den beiden Stichproben



Stripchart, mit "stacking"

Carapaxlängen in den beiden Stichproben



Histogramme/Dichtepolygone und
Stripcharts
geben
ein ausführliches Bild
eines Datensatzes.

Manchmal zu ausführlich.

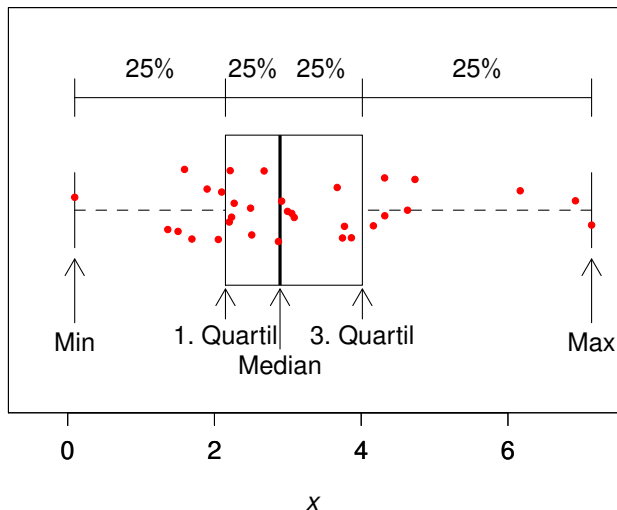
Zu viel Information erschwert den Überblick



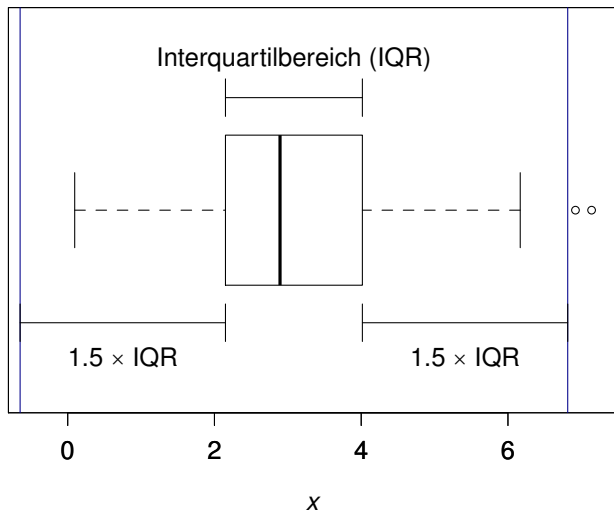
Baum Baum Baum Baum Baum Baum Baum Baum Baum Baum Baum Baum Baum Baum

Wald?

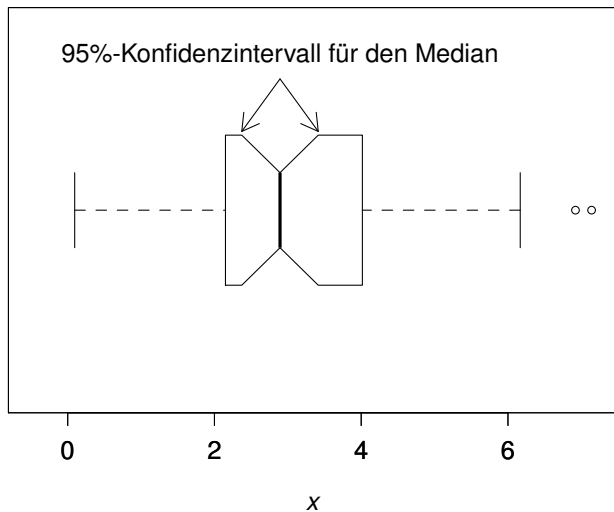
Boxplot, einfache Ausführung



Boxplot, Standard-Ausführung

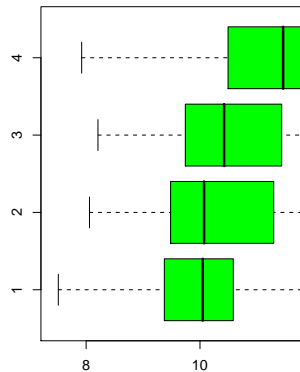
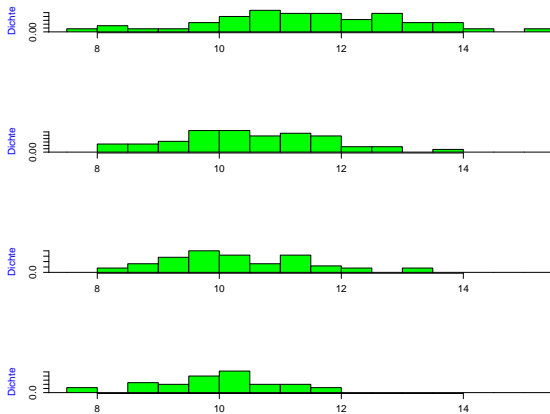


Boxplot, Profi-Ausführung



Beispiel:

Vergleich von mehreren Gruppen



Graphische Trickserien

im Bereich der deskriptiven Statistik / der Kommunikation von numerischen Beobachtungen oder Resultaten:

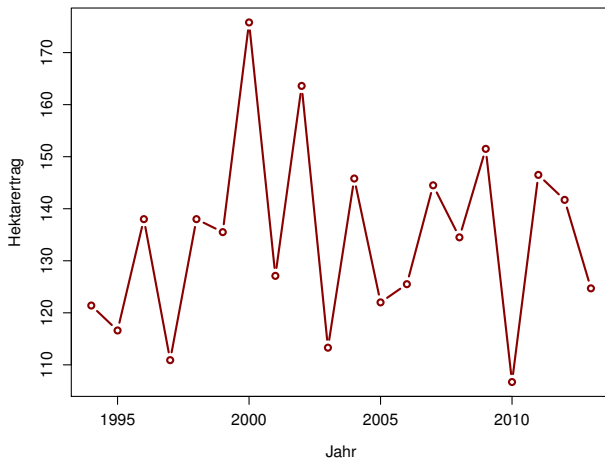
(Graphische) Trickserien / „Aufhübschen“ von Beobachtungen, z.B.

- Irreführende Wahl des Nullpunkts
- Stillschweigende nicht-lineare Transformationen der Achsen
- optische Täuschung durch unpassende 2d/3d-Grafiken
- ...

können den Betrachter (manchmal absichtlich) in die Irre führen.

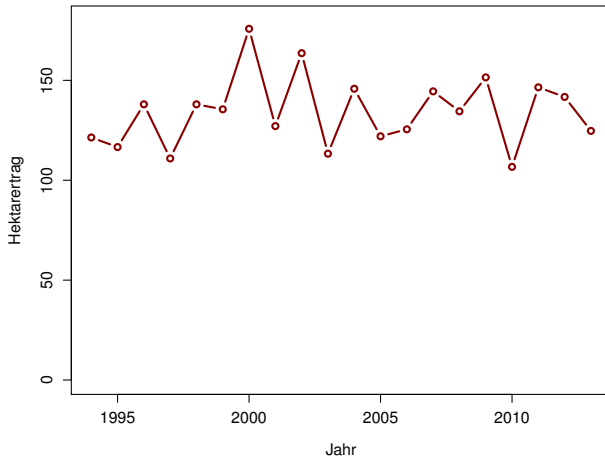
Beunruhigend große Fluktuationen beim Dornfelder?

Hektarerträge Dornfelder, 1994–2013 (in hl)



Beunruhigend große Fluktuationen beim Dornfelder?

Hektarerträge Dornfelder, 1994–2013 (in hl)



Rotwein in RLP: nur ein Tröpfchen?

Bestockte Weinflächen in RLP 2013



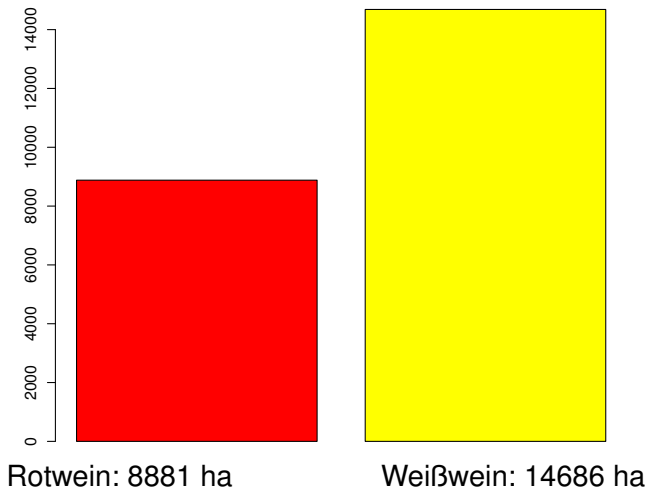
Rotwein: 8881 ha



Weißwein: 14686 ha

Rotwein in RLP: nur ein Tröpfchen?

Bestockte Weinflächen in RLP 2013



Fazit

- 1 Histogramme erlauben einen detaillierten Blick auf die Daten
- 2 Dichtepolygone erlauben Vergleiche zwischen vielen Verteilungen
- 3 Boxplot können große Datenmengen vereinfacht zusammenfassen
- 4 Bei kleinen Datenmengen eher Stripcharts angemessen
- 5 Vorsicht mit Tricks wie 3D oder halbtransparenten Farben

Es ist oft möglich,
das Wesentliche
an einer Stichprobe

mit ein paar Zahlen
zusammenzufassen.

Wesentlich:

1. Wie groß?

Lageparameter

2. Wie variabel?

Streuungsparameter

Eine Möglichkeit
kennen wir schon
aus dem Boxplot:

Lageparameter

Der Median

Streuungsparameter

Der Quartilabstand ($Q_3 - Q_1$)

Der **Median**¹:
die Hälfte der Beobachtungen sind
kleiner,
die Hälfte sind größer.

Der Median ist
das **50%-Quantil**
der Daten.

¹„saloppe“ Definition (wir sehen gleich die präzise Definition)

Die Quartile

Das erste Quartil², Q_1 :
ein Viertel der Beobachtungen
sind kleiner,
drei Viertel sind größer.

Q_1 ist das
25%-Quantil
der Daten.

²„saloppe“ Definition (wir sehen gleich die präzise Definition)

Die Quartile

Das dritte Quartil³, Q_3 :
drei Viertel der Beobachtungen
sind kleiner,
ein Viertel sind größer.

Q_3 ist das
75%-Quantil
der Daten.

³„saloppe“ Definition (wir sehen gleich die präzise Definition)

(Empirische) Quantile, allgemein

Seien n (reelle) Beobachtungswerte x_1, x_2, \dots, x_n gegeben, $\alpha \in (0, 1)$.

q ist (ein) α -Quantil der n Beobachtungswerte, wenn gilt

$$\frac{1}{n} |\{1 \leq i \leq n : x_i \leq q\}| \geq \alpha \text{ und } \frac{1}{n} |\{1 \leq i \leq n : x_i \geq q\}| \geq 1 - \alpha.$$

Bem.: Im Allgemeinen ist ein α -Quantil nicht eindeutig:

Seien $x_{(1)} \leq x_{(2)} \leq \dots \leq x_{(n)}$ die der Größe nach sortierten Werte.

Wenn $\alpha = \frac{k}{n}$ mit $1 \leq k < n$, so ist jeder Wert $q \in [x_{(k)}, x_{(k+1)}]$ ein α -Quantil,

denn $|\{i : x_i \leq x_{(k)}\}| \geq k$, $|\{i : x_i \geq x_{(k)}\}| \geq n - k + 1$.

Wenn $n\alpha \notin \{1, \dots, n-1\}$, so ist das α -Quantil der Wert $x_{(k)}$ mit $k = \lceil \alpha n \rceil$.

(Empirische) Quantile, allgemein II

n (reelle) Beobachtungswerte x_1, x_2, \dots, x_n gegeben,
 $\alpha \in (0, 1)$.

(ein) α -Quantil q der n Beobachtungswerte erfüllt

$$\frac{1}{n} |\{1 \leq i \leq n : x_i \leq q\}| \geq \alpha \text{ und } \frac{1}{n} |\{1 \leq i \leq n : x_i \geq q\}| \geq 1 - \alpha.$$

Bem.:

- Die Definition passt zu unserer früheren Definition für Verteilungen, wenn man die *empirische Verteilung* $\frac{1}{n} \sum_{i=1}^n \delta_{x_i}$ betrachtet.
- In der Literatur (und auch in Statistik-Software) sind verschiedene Interpolationen üblich, um „das“ α -Quantil stetig in α zu machen.

(In R siehe etwa `help(quantile)`, es sind 9 Varianten implementiert.)

n (reelle) Beobachtungswerte x_1, x_2, \dots, x_n

Am häufigsten werden benutzt:

Lageparameter

Der **Mittelwert** $\bar{x} := \frac{1}{n} \sum_{i=1}^n x_i$

Streuungsparameter

Die **Standardabweichung** s (bzw. σ)

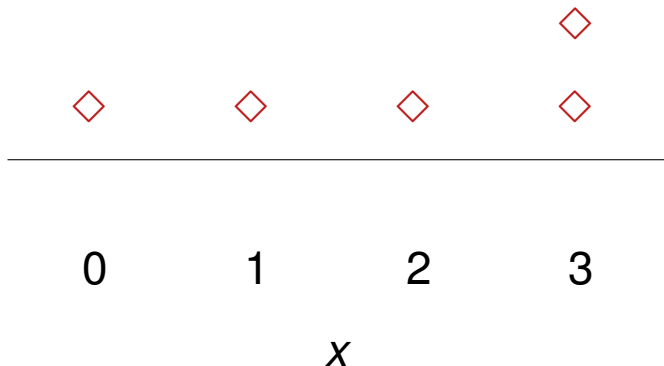
wobei

$\sigma^2 = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2$ die (empirische) Varianz

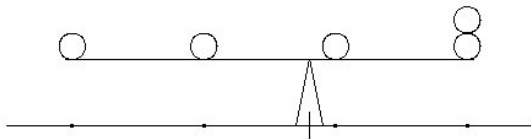
$s^2 = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2$ die korrigierte Stichproben-Varianz
 (= $\frac{n}{n-1} \sigma^2$)

Erinnerung: Geometrische Bedeutung des Mittelwerts Der Schwerpunkt

Wo muß der Drehpunkt sein, damit die Waage im Gleichgewicht ist?



$$m = 1,8 ?$$



richtig

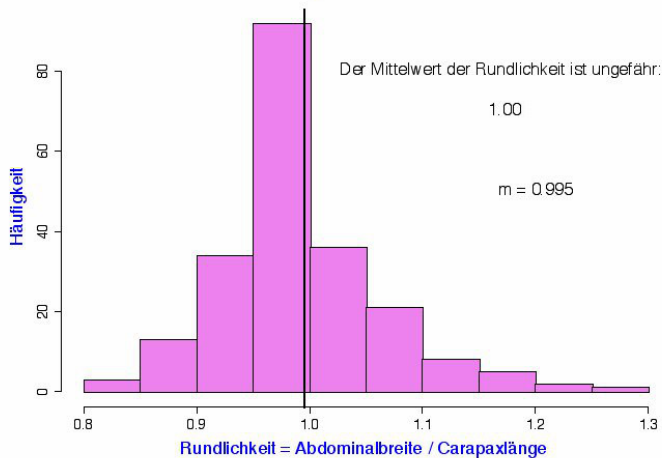
Oft kann man „mit dem bloßen Auge“
anhand eines Histogramms den
Mittelwert gut einschätzen.

Beispiel: *Galathea intermedia*

„Rundlichkeit“

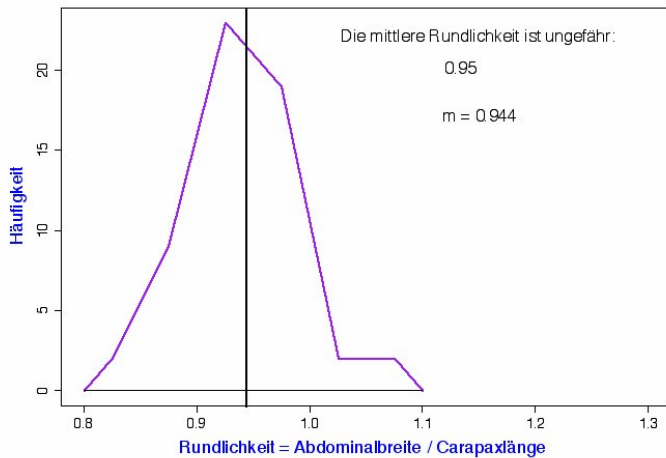
:=

Abdominalbreite / Carapaxlänge

Nichteiertragende Weibchen 6.9.88

Beispiel:

3.11.88

Nichteiertragende Weibchen 3. 11. 88

Die Standardabweichung (auch: Streuung)

Wie weit weicht
eine typische Beobachtung
vom
Mittelwert
ab ?

Mit n oder $n - 1$ berechnen?

Die Standardabweichung σ eines Zufallsexperiments mit n gleichwahrscheinlichen Ausgängen x_1, \dots, x_n (z.B. Würfelwurf) ist definiert durch (vgl. Def. 1.76)

$$\sigma = \sqrt{\frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2}.$$

Wenn es sich bei x_1, \dots, x_n um Beobachtungswerte in einer Stichprobe handelt, verwendet man eher

$$s = \sqrt{\frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2}.$$

s als Schätzer für σ

Wir werden sehen:

Wenn X_1, \dots, X_n u.i.v. *Zufallsvariablen* mit Varianz $\text{Var}[X_1] = \sigma^2$,

$$\bar{X} := \frac{1}{n} \sum_{i=1}^n X_i,$$

so hat die *Zufallsvariable*

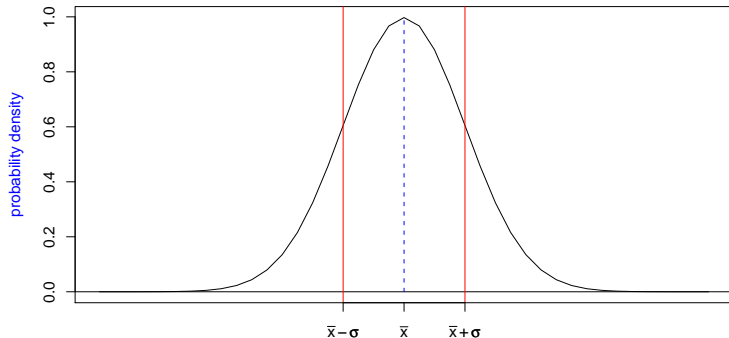
$$S^2 := \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X})^2$$

die Eigenschaft

$$\mathbb{E}[S^2] = \sigma^2.$$

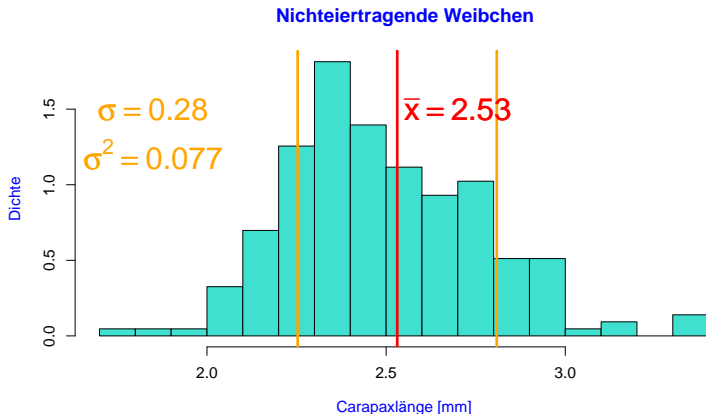
Faustregel für die Standardabweichung

Bei ungefähr glockenförmigen (also eingipfligen und symmetrischen) Verteilungen liegen ca. 2/3 der Verteilung zwischen $\bar{x} - \sigma$ und $\bar{x} + \sigma$.



Oft kann man so die Standardabweichung „mit bloßem Auge“ abschätzen.

Standardabweichung der Carapaxlängen nichtiertragender Weibchen vom 6.9.88



Hier liegt der Anteil zwischen $\bar{x} - \sigma$ und $\bar{x} + \sigma$ bei 72%.

Übrigens: Einschlägige R-Befehle

Mittelwert (`mean`), Standardabweichung (`sd`), Median, und Quantile

```
mean(x)
sd(x)
median(x)
quantile(x, 0.25, type=1)
quantile(x, 0.75, type=1)
summary(x)
```

Boxplot, Histogramm

```
boxplot(x)
hist(x)      (für Dichtehistogramm: hist(x, prob=T))
```


Ein Dichtepolygon gewinnt man z.B. via

```
h <- hist(x)
plot(h$mids, h$density, type='l')
```


Mittelwert und Standardabweichung. . .

- charakterisieren die Daten gut, falls deren Verteilung (zumindest in etwa) glockenförmig ist
- und müssen andernfalls mit Vorsicht interpretiert werden.

Wir betrachten dazu einige Lehrbuch-Beispiele aus der Biologie, siehe z.B.

 M. Begon, C. R. Townsend, and J. L. Harper.
Ecology: From Individuals to Ecosystems.
Blackell Publishing, 4 edition, 2008.

(Wir verwenden an die Originalpublikationen angelehnte simulierte Daten, nehmen Sie also nicht alle Datenpunkte wörtlich.)

Bachstelzen fressen Dungfliegen

Räuber



Bachstelze (White Wagtail)
Motacilla alba alba

image (c) by Artur Mikolajewski

Beute



Gelbe Dungfliege
Scatophaga stercoraria

image (c) by Viatour Luc

Vermutung

- Die Fliegen sind unterschiedlich groß
- Effizienz für die Bachstelze = $\text{Energiegewinn} / \text{Zeit}$ zum Fangen und fressen
- Laborexperimente lassen vermuten, dass die Effizienz bei 7mm großen Fliegen maximal ist.

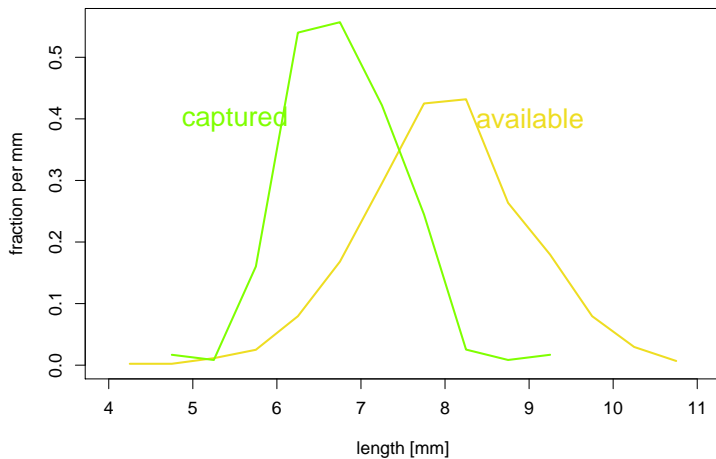


N.B. Davies.

Prey selection and social behaviour in wagtails (Aves: Motacillidae).

J. Anim. Ecol., 46:37–57, 1977.

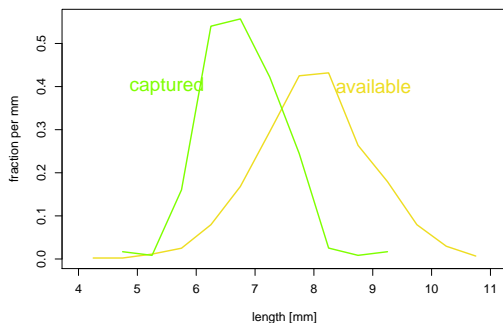
dung flies: available, captured



Vergleich der Größenverteilungen

	captured		available
Mittelwert	6.79	<	7.99
Standardabweichung	0.69	<	0.96

dung flies: available, captured



Interpretation

Die Bachstelzen bevorzugen Dungfliegen, die etwa 7mm groß sind.

Hier waren die Verteilungen glockenförmig und es genügte 4 Werte (die beiden Mittelwerte und die beiden Standardabweichungen), um die Daten adäquat zu beschreiben.



Nephila madagascariensis

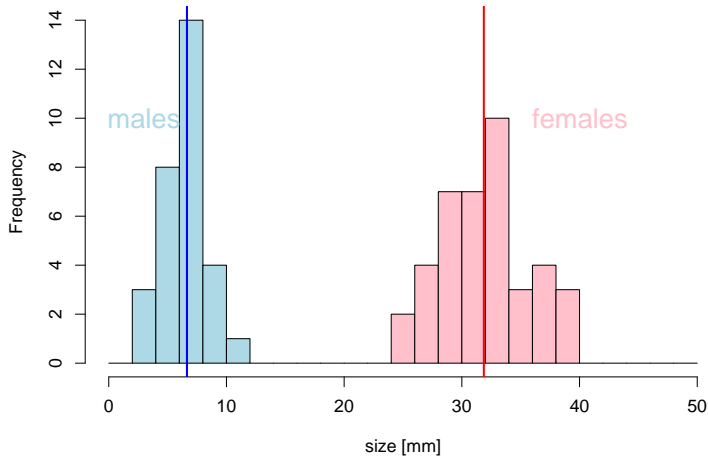
image (c) by Bernard Gagnon

Simulierte Daten:

Eine Stichprobe von 70 Spinnen

Mittlere Größe: 21,06 mm

Standardabweichung der Größe: 12,94 mm

***Nephila madagascariensis* (n=70)**



Nephila madagascariensis

image (c) by Arthur Chapman

Fazit des Spinnenbeispiels

Wenn die Daten aus verschiedenen Gruppen zusammengesetzt sind, die sich bezüglich des Merkmals deutlich unterscheiden, kann es sinnvoll sein, Kenngrößen wie den Mittelwert für jede Gruppe einzeln zu berechnen.

Kupfertolerantes Rotes Straußgras



Rotes Straußgras
Agrostis tenuis

image (c) Kristian Peters



Kupfer
Cuprum

Hendrick met de Bles

Anpassung an Kupfer?

- Pflanzen, denen das Kupfer schadet, haben kürzere Wurzeln.
- Die Wurzellängen von Pflanzen aus der Umgebung von Kupferminen wird gemessen.
- Samen von unbelasteten Wiesen werden bei Kupferminen eingesät.
- Die Wurzellängen dieser “Wiesenpflanzen” werden gemessen.



A.D. Bradshaw.

Population Differentiation in *agrostis tenuis* Sibth. III. populations in varied environments.

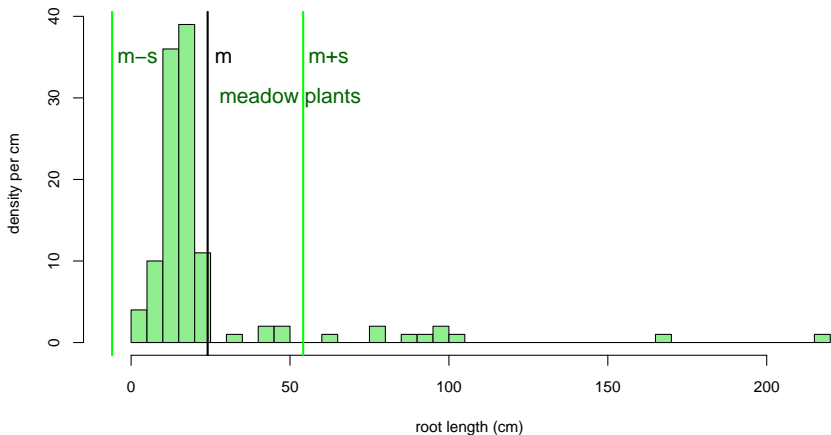
New Phytologist, 59(1):92 – 103, 1960.



T. McNeilly and A.D Bradshaw.

Evolutionary Processes in Populations of Copper Tolerant *Agrostis tenuis* Sibth.

Browntop Bent (n=50)

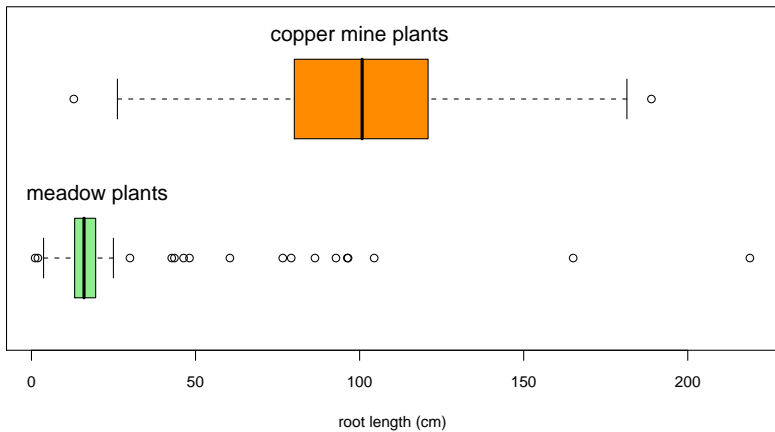


2/3 der Wurzellängen innerhalb $[m-sd, m+sd]$???? **Nein!**

Fazit des Straußgras-Beispiels

Manche Verteilungen können nur mit mehr als zwei Variablen angemessen beschrieben werden.

z.B. mit den fünf Werten der Boxplots:
min, Q_1 , median, Q_3 , max

Browntop Bent n=50+50

Schlussfolgerung

Viele Datenverteilungen sind annähernd glockenförmig und können durch den **Mittelwert** und die **Standardabweichung** hinreichend beschrieben werden.

Es gibt aber auch Ausnahmen. Also: **Besser** ist es, die Daten auch graphisch zu untersuchen, und sich **nicht** allein auf numerische Kenngrößen zu verlassen.