# Einführung in die Stochastik, WS 2023/24

# 6. Statistik

# 6.1 Deskriptive Statistik

Matthias Birkner

http://www.staff.uni-mainz.de/birkner/GrundlStoch_2324/

# Inhalt

Viele Menschen stehen „Statistik“ kritisch gegenüber:

*It is easy to lie with statistics.*

*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.

# Worum geht es in der Statistik?

Die Welt ist voller Variabilität.

Wie geht man mit variablen Daten um?

# 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

# 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 spezifiert werden.

# 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 spezifiert 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?

# 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

# Inhalt

# Ein Beispiel

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

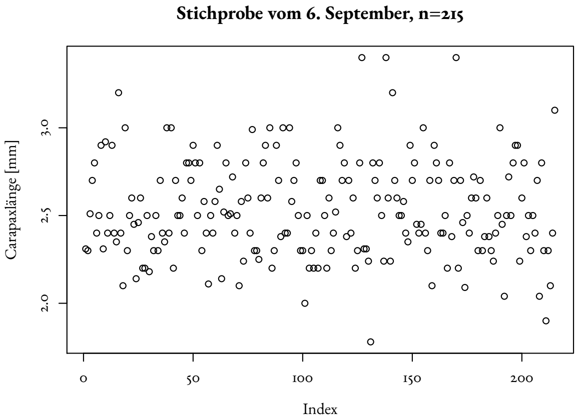
# Ein Beispiel

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 ($n = 215$)

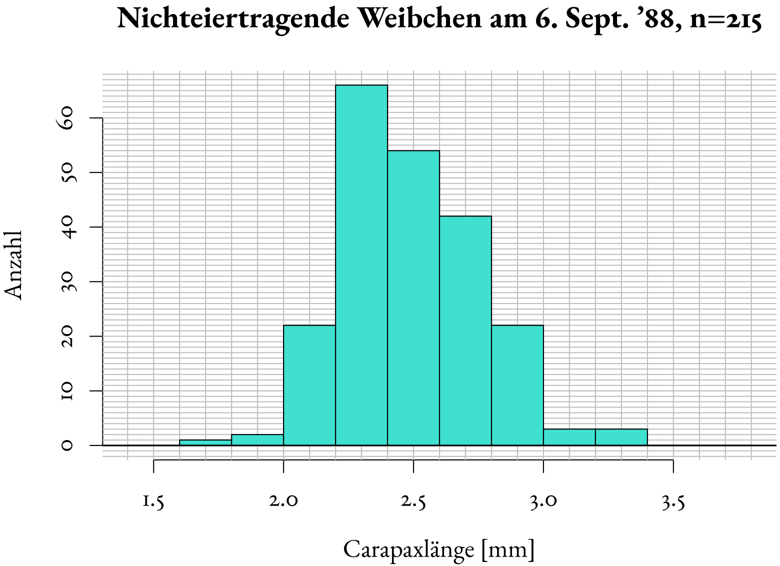
| | | | | | | | |
|---|---|---|---|---|---|---|---|
| 2,9 | 3,0 | 2,9 | 2,5 | 2,7 | 2,9 | 2,9 | 3,0 |
| 3,0 | 2,9 | 3,4 | 2,8 | 2,9 | 2,8 | 2,8 | 2,4 |
| 2,8 | 2,5 | 2,7 | 3,0 | 2,9 | 3,2 | 3,1 | 3,0 |
| 2,7 | 2,5 | 3,0 | 2,8 | 2,8 | 2,8 | 2,7 | 3,0 |
| 2,6 | 3,0 | 2,9 | 2,8 | 2,9 | 2,9 | 2,3 | 2,7 |
| 2,6 | 2,7 | 2,5 | . | . | . | . | . |

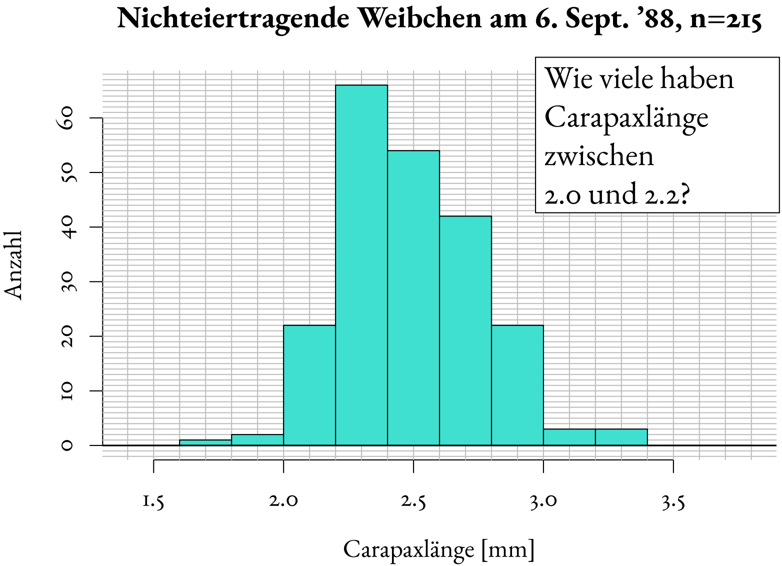

Stichprobe vom 6. September, n=215
Carapaxlänge [mm]
3.0
2.5
2.0
0
50
100
150
200
Index

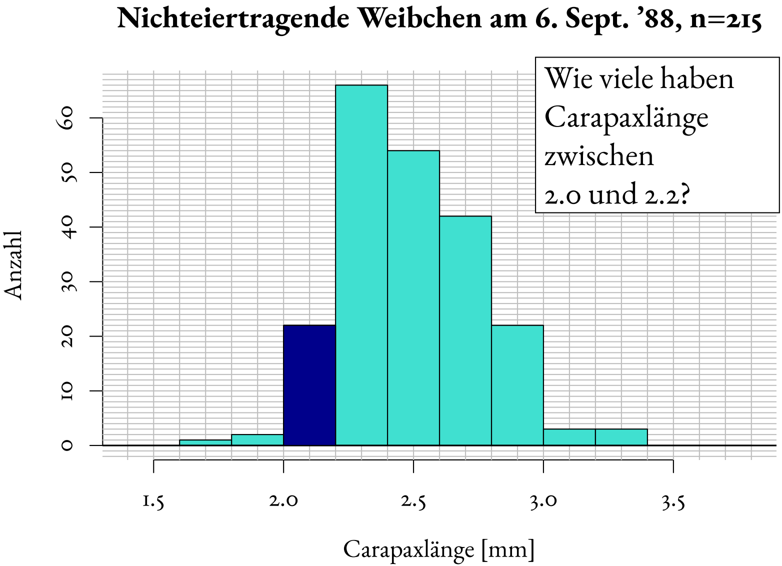
# Inhalt

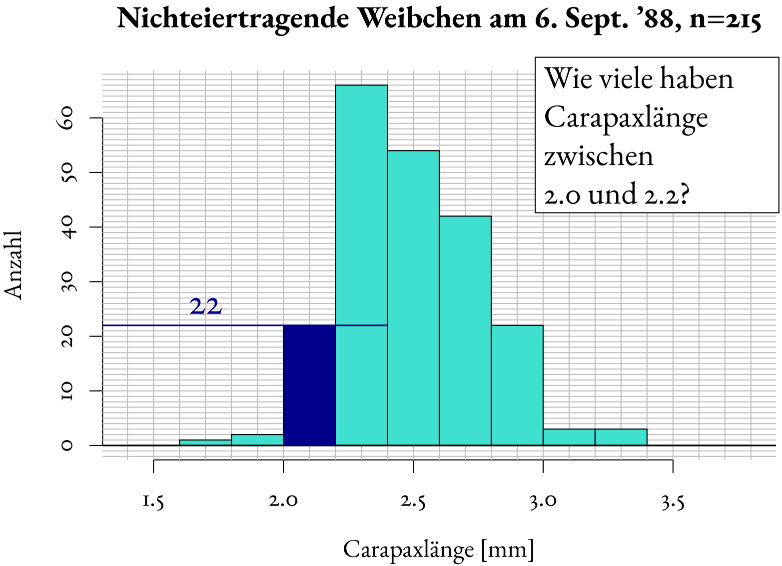
# Eine Möglichkeit der graphischen Darstellung:

# das Histogramm


Nichteiertragende Weibchen am 6. Sept. '88, n=215
Anzahl
0
10
20
30
40
50
60
1.5
2.0
2.5
3.0
3.5
Carapaxlänge [mm]

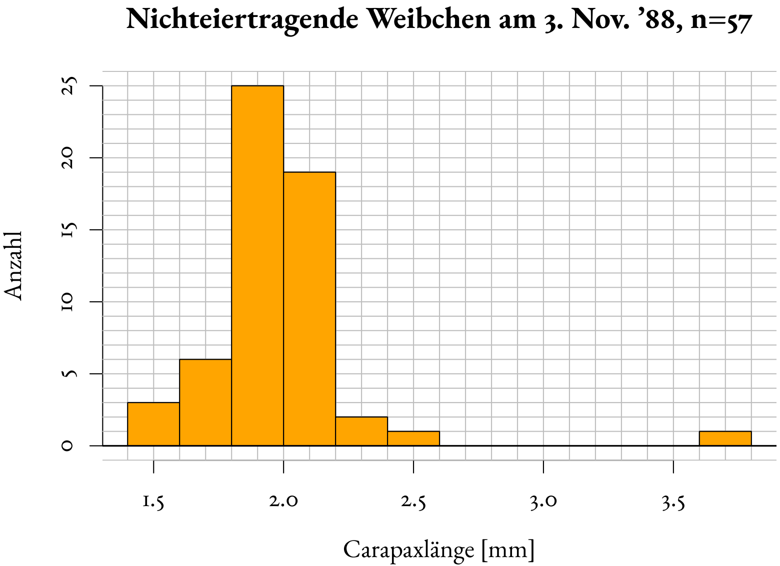

Nichteiertragende Weibchen am 6. Sept. ’88, n=215
Wie viele haben Carapaxlänge zwischen 2.0 und 2.2?
Anzahl
0
10
20
30
40
50
60
1.5
2.0
2.5
3.0
3.5
Carapaxlänge [mm]


Nichteiertragende Weibchen am 6. Sept. '88, n=215
Wie viele haben Carapaxlänge zwischen 2.0 und 2.2?
Anzahl
0
10
20
30
40
50
60
1.5
2.0
2.5
3.0
3.5
Carapaxlänge [mm]

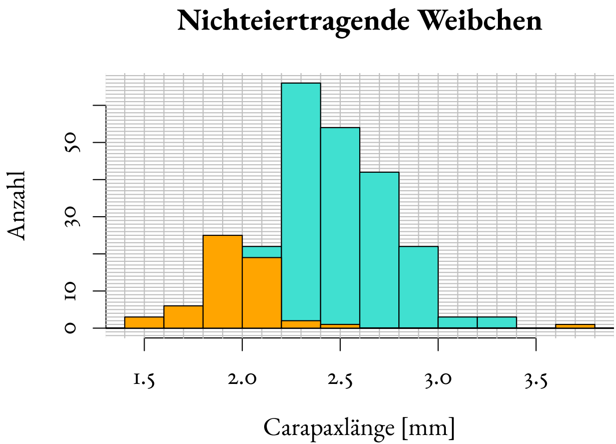

Nichteiertragende Weibchen am 6. Sept. ’88, n=215
Wie viele haben Carapaxlänge zwischen 2.0 und 2.2?
22
Anzahl
0
10
20
30
40
50
60
1.5
2.0
2.5
3.0
3.5
Carapaxlänge [mm]

Analoge Daten zwei Monate später (3.11.88):

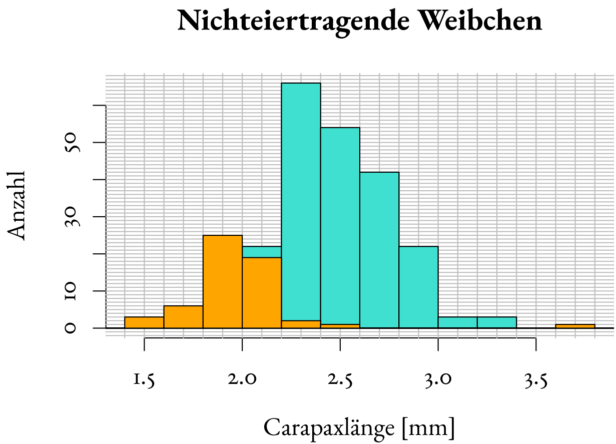
Analoge Daten zwei Monate später (3.11.88):



# Vergleich der beiden Verteilungen:



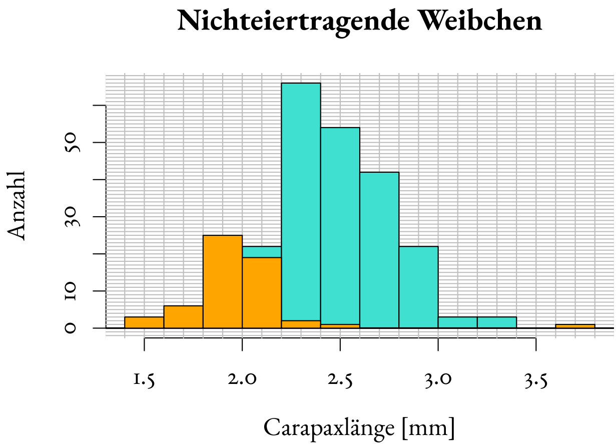
Vergleich der beiden Verteilungen:




Problem: ungleiche Stichprobenumfänge:

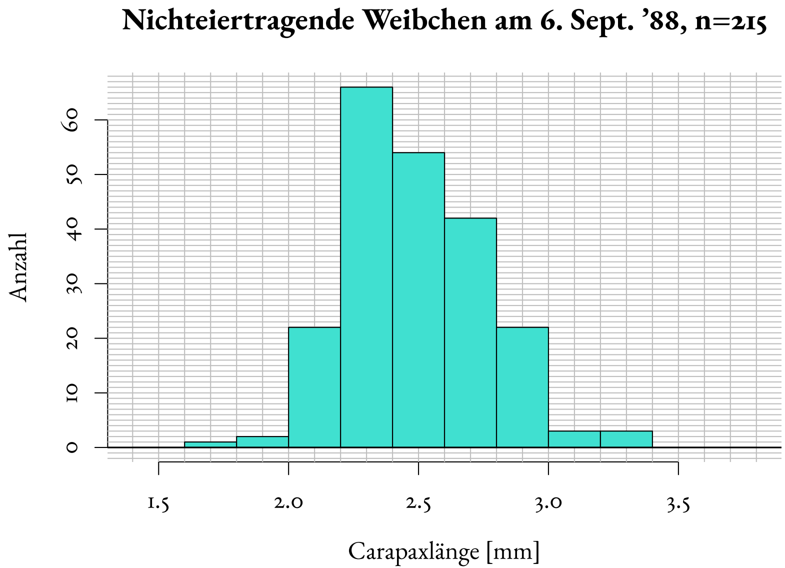
6.Sept: $n = 215$
3.Nov : $n = 57$

Vergleich der beiden Verteilungen:

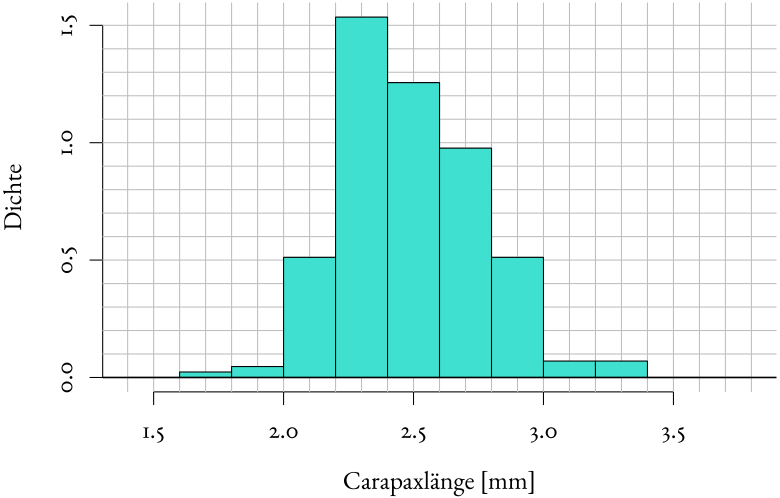



Problem: ungleiche Stichprobenumfänge: 6.Sept: $n = 215$
3.Nov : $n = 57$

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

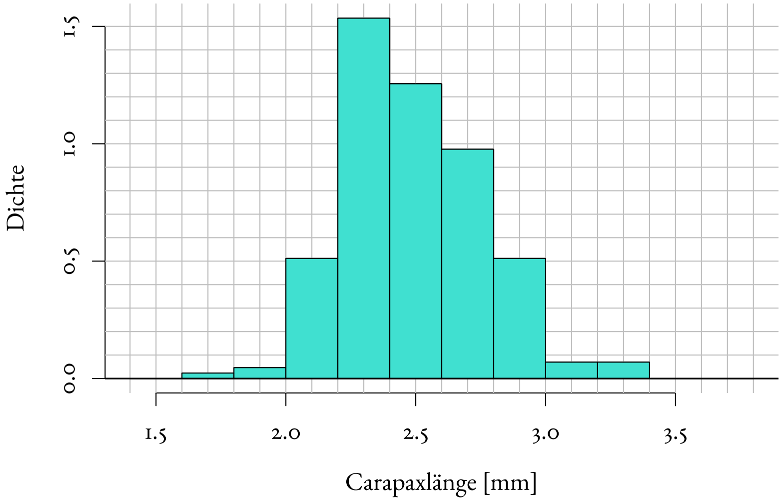

Nichteiertragende Weibchen am 6. Sept. ’88, n=215
Anzahl
0
10
20
30
40
50
60
1.5
2.0
2.5
3.0
3.5
Carapaxlänge [mm]

## Nichteiertragende Weibchen am 6. Sept. '88, n=215

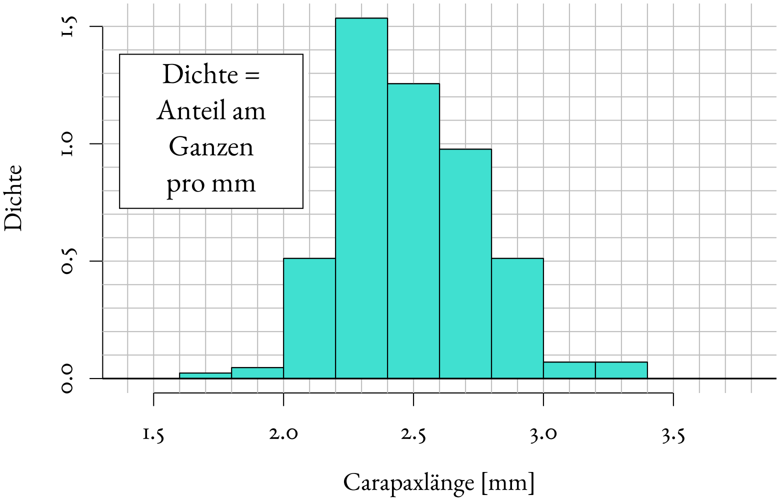


Die neue
vertikale Koordinate
ist jetzt eine
Dichte
(engl. density).

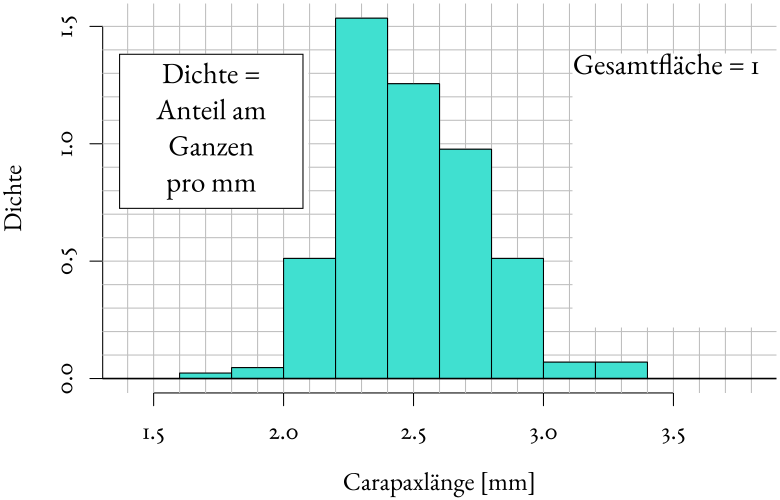
# Nichteiertragende Weibchen am 6. Sept. '88, n=215



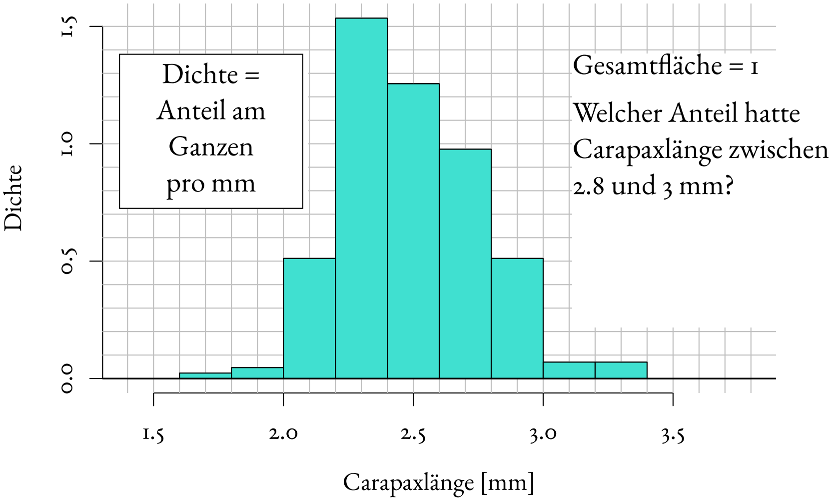
# Nichteiertragende Weibchen am 6. Sept. '88, n=215



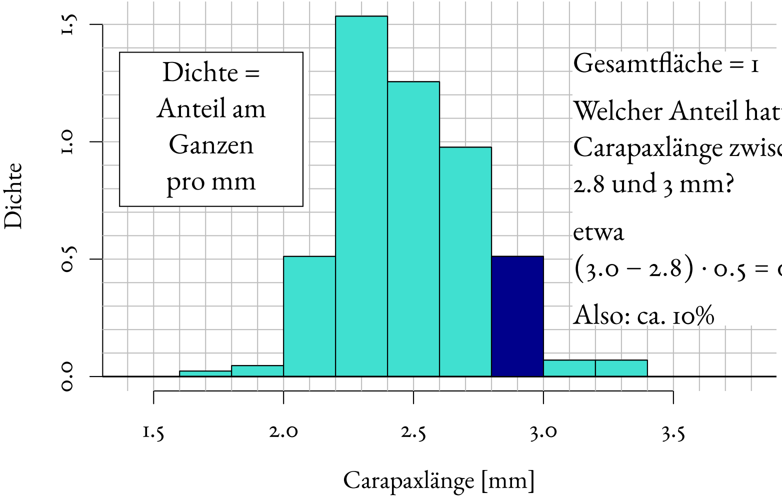
# Nichteiertragende Weibchen am 6. Sept. ’88, n=215



# Nichteiertragende Weibchen am 6. Sept. ’88, n=215



# Nichteiertragende Weibchen am 6. Sept. ’88, n=215




Gesamtfläche = 1

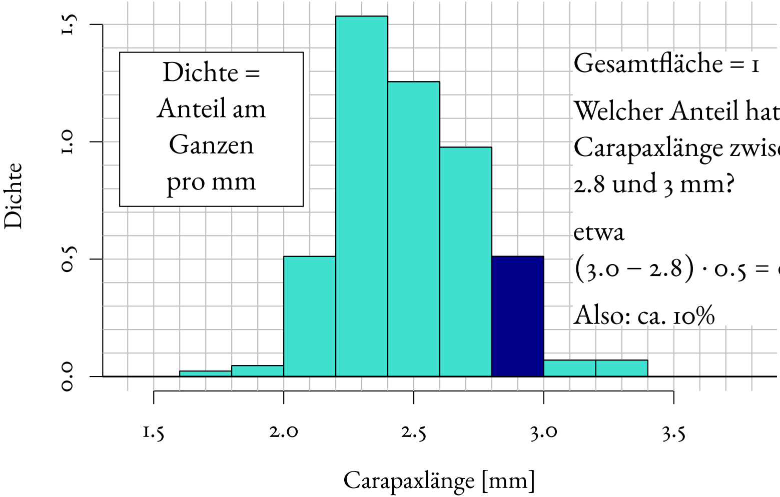
Welcher Anteil hatte Carapaxlänge zwischen 2.8 und 3 mm?

etwa

$(3.0 - 2.8) \cdot 0.5 = 0.1$

Also: ca. 10%

# Nichteiertragende Weibchen am 6. Sept. ’88, n=215




Gesamtfläche = 1

Welcher Anteil hatte Carapaxlänge zwischen 2.8 und 3 mm?

etwa

$(3.0 - 2.8) \cdot 0.5 = 0.1$

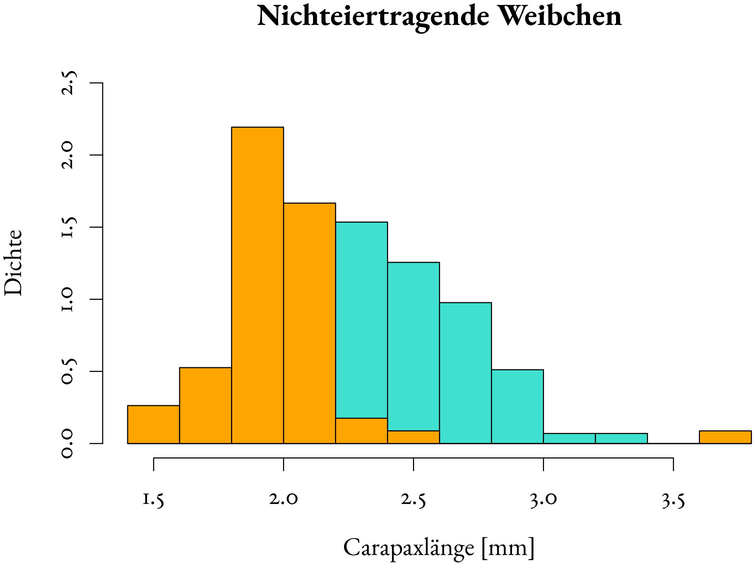
Also: ca. 10%

(Ganz penibel: Ich schreibe „ca. 10%“, denn die exakte Balkenhöhe ist 0.5116..., aber so genau wollen wir es hier gar nicht angeben.)

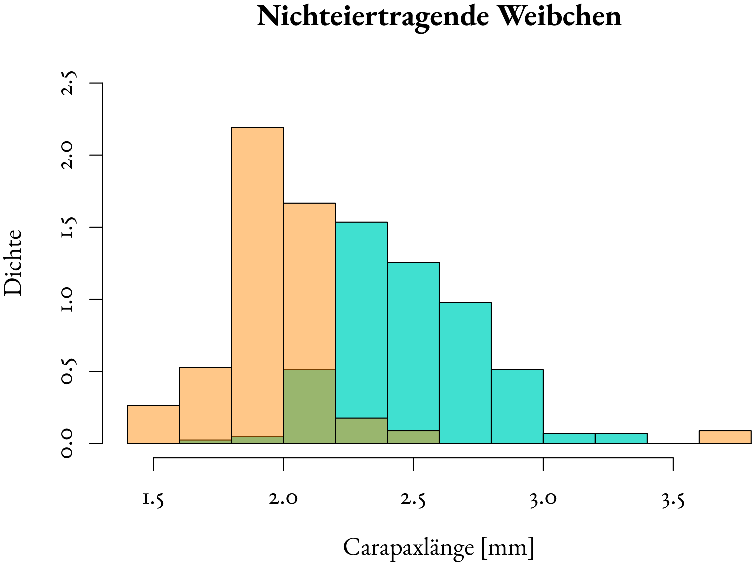
# Die beiden Histogramme sind jetzt vergleichbar

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

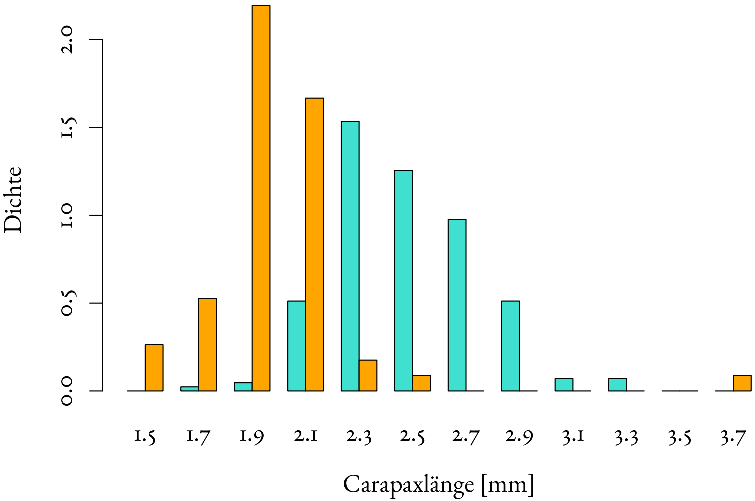
Versuche, die Histogramme zusammen zu zeigen:



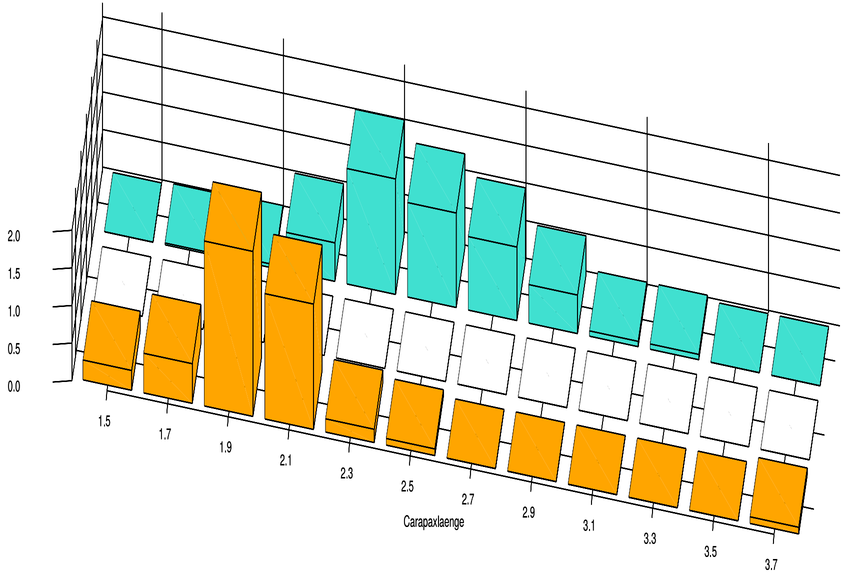
Versuche, die Histogramme zusammen zu zeigen:



# Versuche, die Histogramme zusammen zu zeigen:



# Versuche, die Histogramme zusammen zu zeigen:



# Vorschlag

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

# Vorschlag

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

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

# Problem

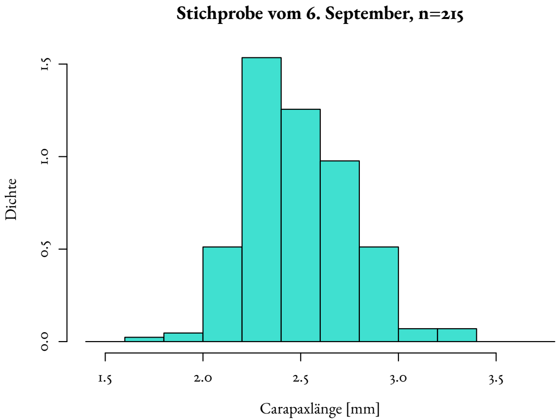
Histogramme kann man nicht ohne weiteres
in demselben Graphen
darstellen,

# Problem

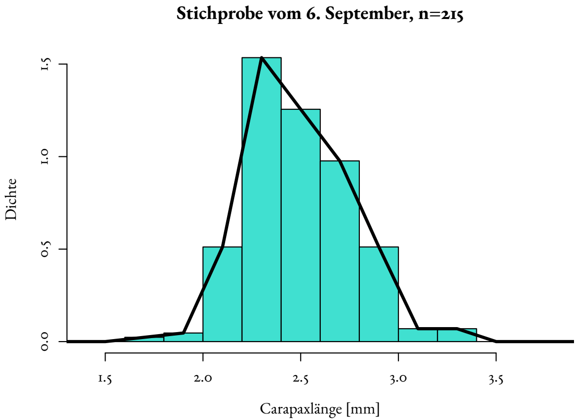
Histogramme kann man nicht ohne weiteres
in demselben Graphen
darstellen,

weil sie einander
überdecken würden.

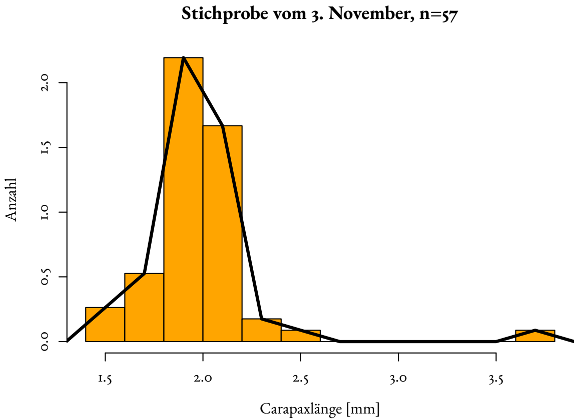
# Einfache und klare Lösung: Dichtepolygone



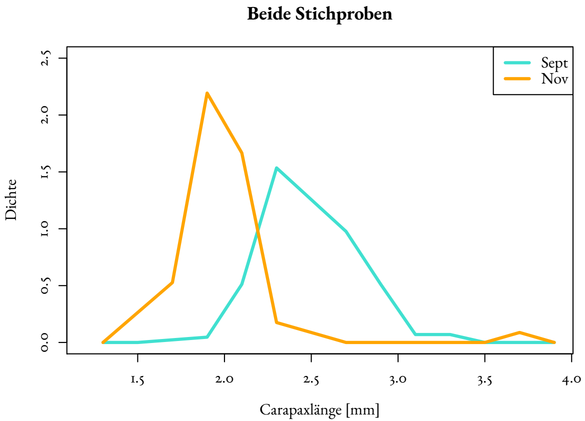
# Einfache und klare Lösung: Dichtepolygone



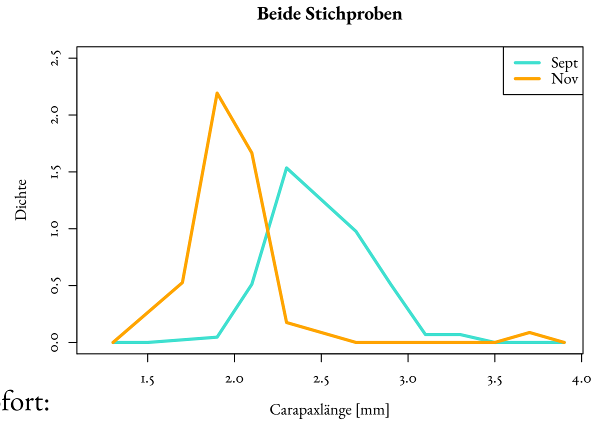
# Einfache und klare Lösung: Dichtepolygone



# Zwei (oder mehr) Dichtepolygone in einem Plot



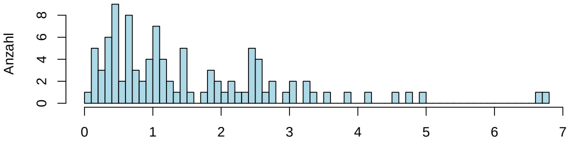
# Zwei (oder mehr) Dichtepolygone in einem Plot



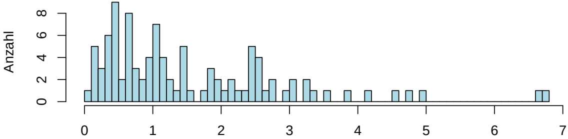

Man sieht sofort:

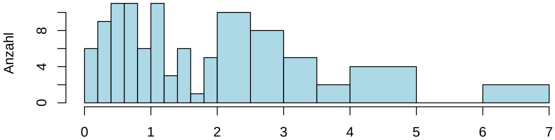
Die Verteilung in der Stichprobe vom November ist gegenüber der vom September nach links verschoben (und sie ist auch stärker um den häufigsten Wert konzentriert).

# Anzahl vs. Dichte

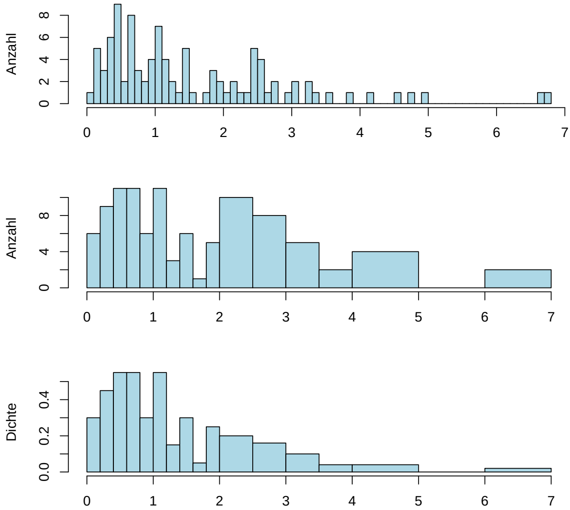


# Anzahl vs. Dichte

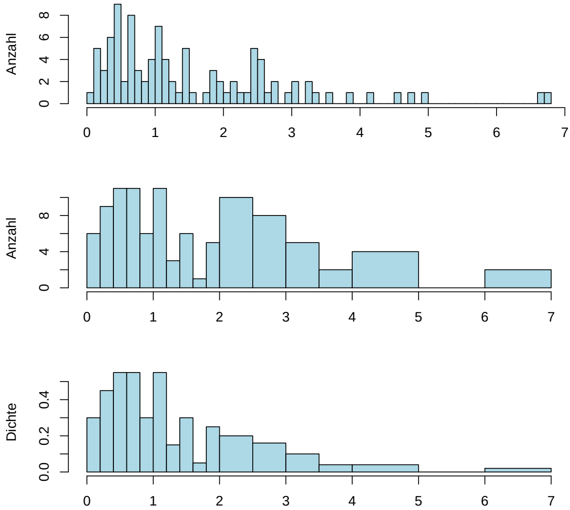


# Anzahl vs. Dichte



# Anzahl vs. Dichte

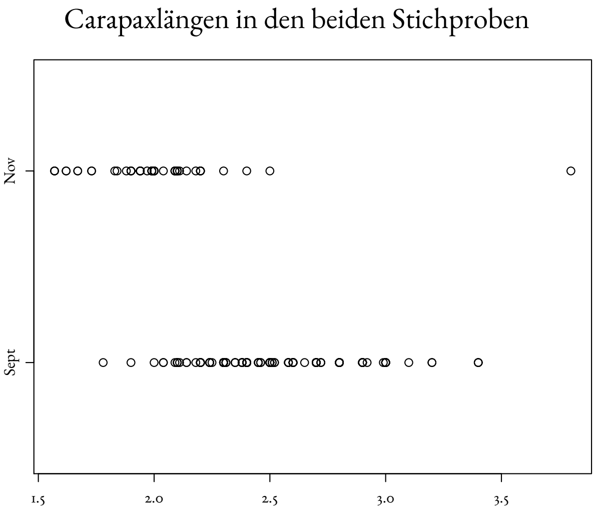



Also:

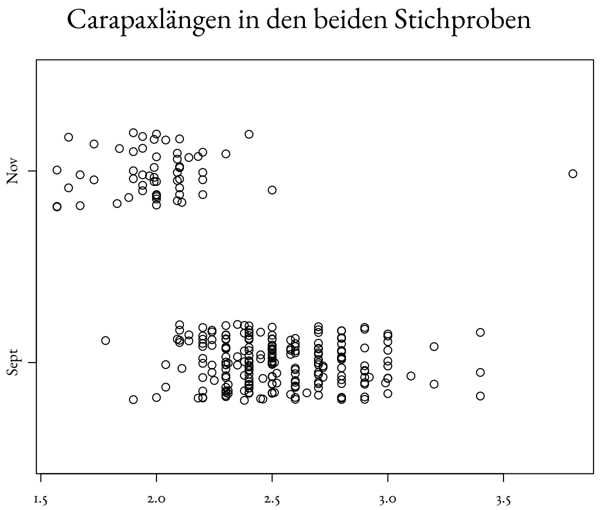
Bei Histogrammen mit ungleichmäßiger Unterteilung immer Dichten verwenden!

# Inhalt

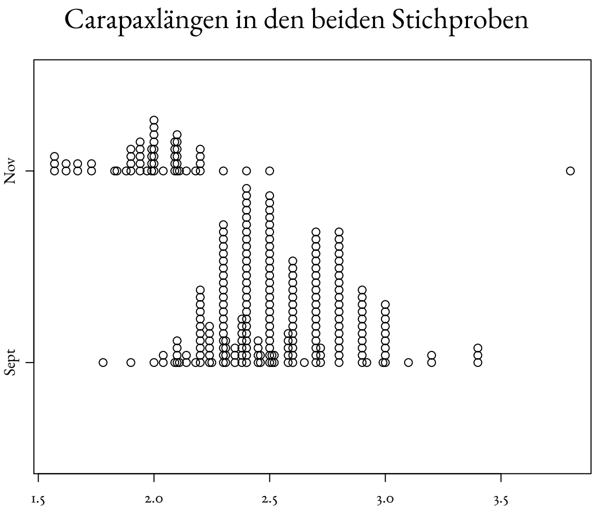
# Stripchart, einfach



# Stripchart, mit “jitter”



# Stripchart, mit “stacking”



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

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

Manchmal zu ausführlich.

# Inhalt

# Zu viel Information erschwert den Überblick

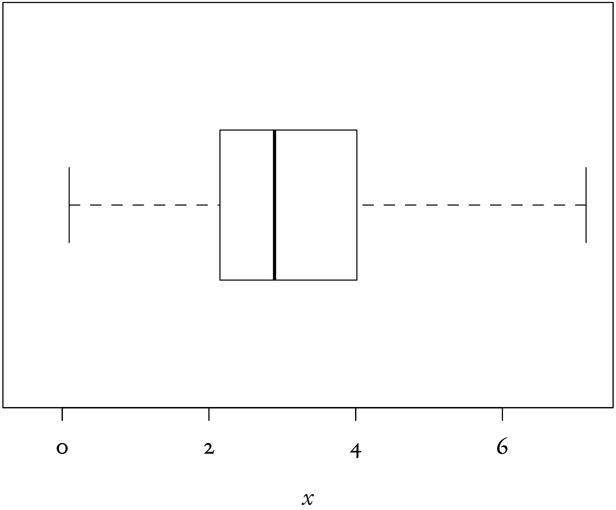
Baum Baum Baum Baum Baum Baum Baum Baum Baum Baum Baum Baum Baum

# Zu viel Information erschwert den Überblick

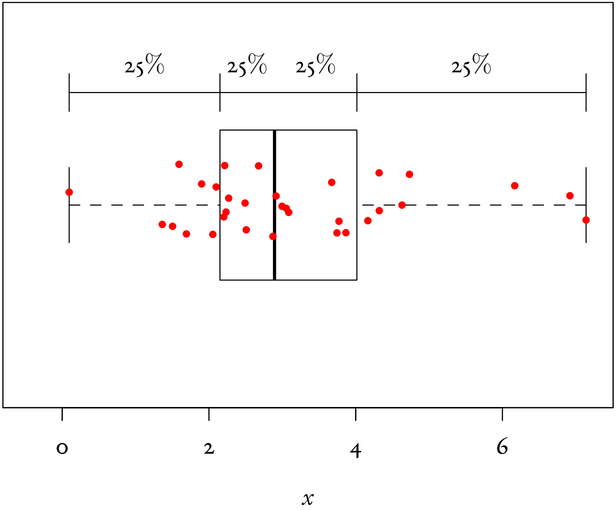
Baum Baum Baum Baum Baum Baum Baum Baum Baum Baum Baum Baum Baum

Wald?

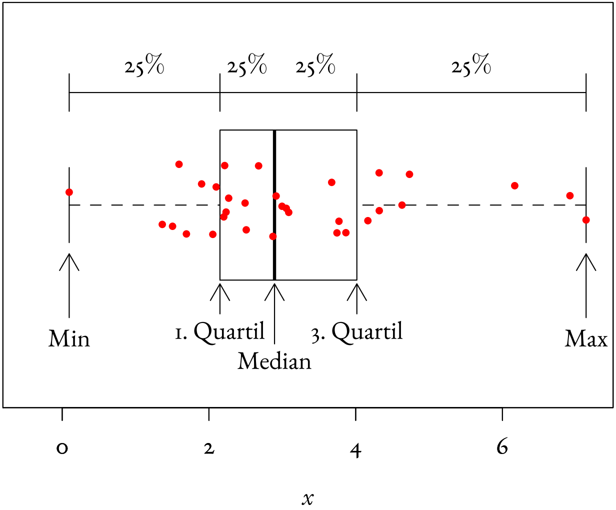
# Boxplot, einfache Ausführung



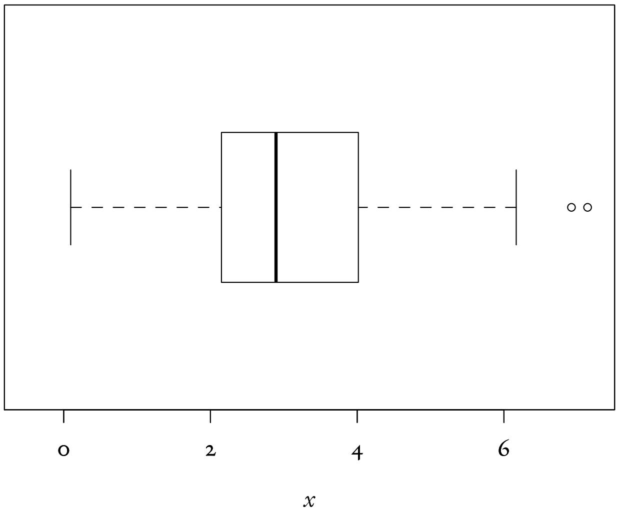
# Boxplot, einfache Ausführung



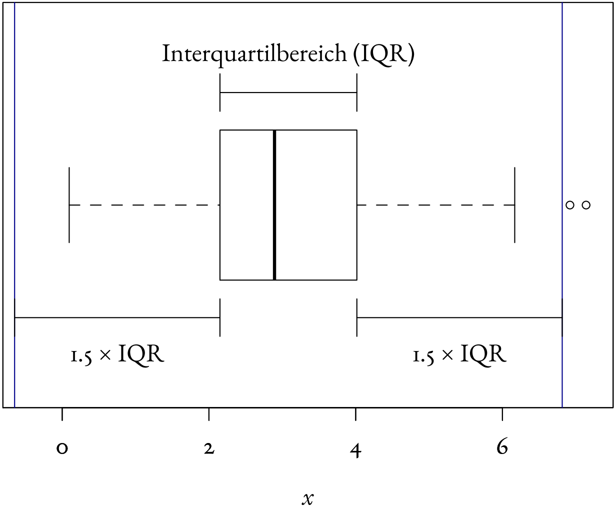
# Boxplot, einfache Ausführung



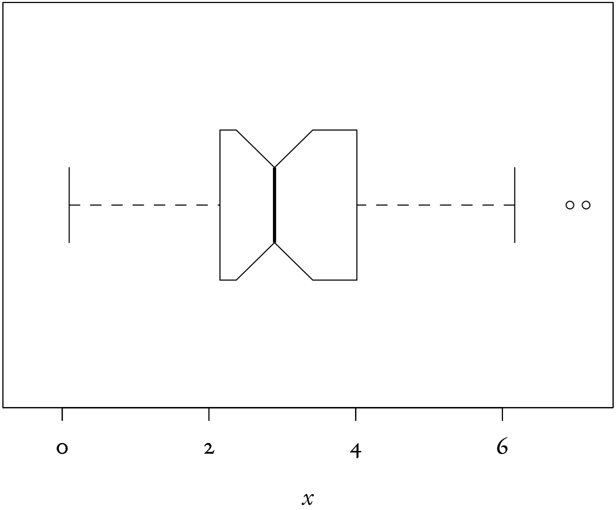
# Boxplot, Standard-Ausführung



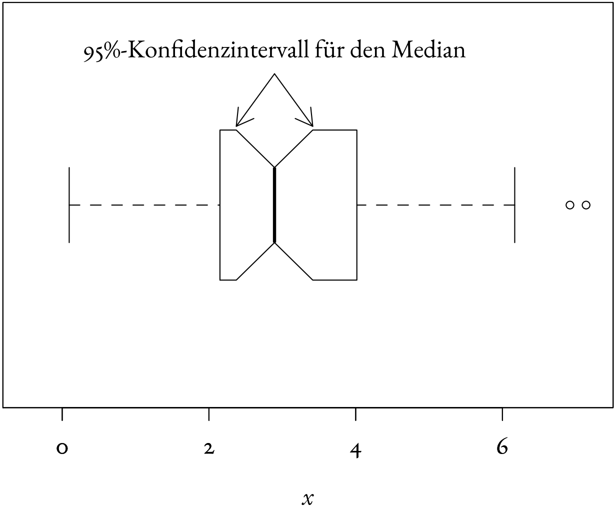
# Boxplot, Standard-Ausführung



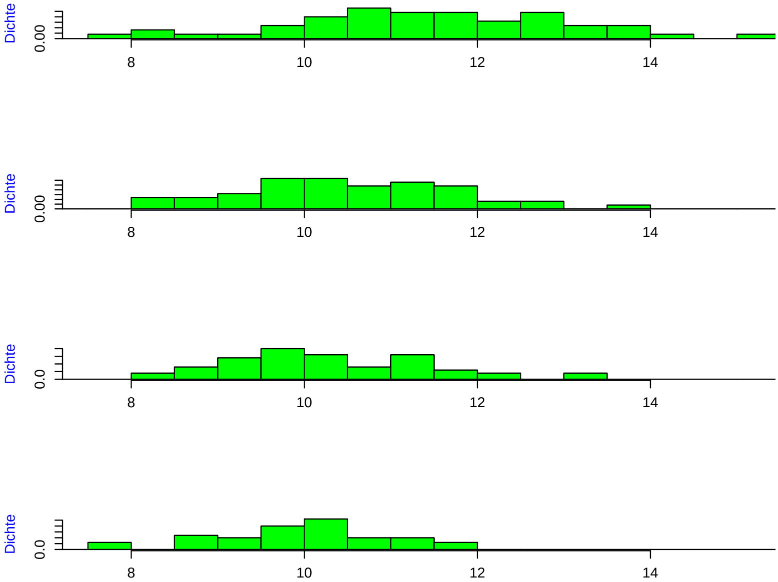
# Boxplot, Profi-Ausführung

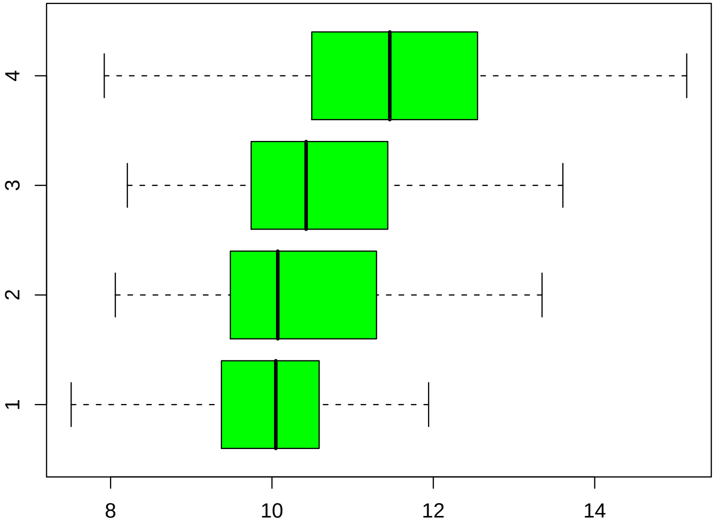


# Boxplot, Profi-Ausführung



# Beispiel: Vergleich von mehreren Gruppen


Dichte
0.00
8
10
12
14
Dichte
0.00
8
10
12
14
Dichte
0.0
8
10
12
14
Dichte
0.0
8
10
12
14


4
3
2
1
8
10
12
14

# Inhalt

# Graphische Tricksereien

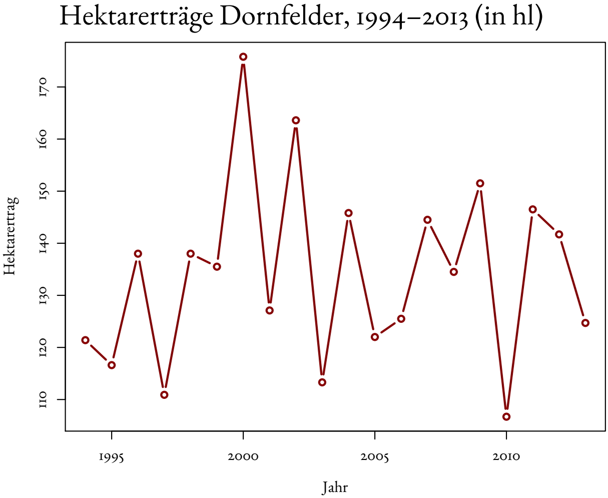
im Bereich der deskriptiven Statistik / der Kommunikation von numerischen Beobachtungen oder Resultaten:

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

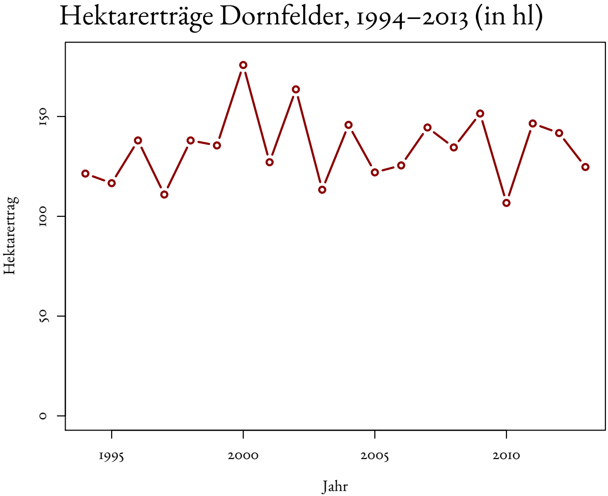
- Irreführende Wahl des Nullpunkts
- Stillschweigende nicht-lineare Transformationen der Achsen
- optische Täuschung durch unpassende 2d/3d-Grafiken
- ...

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

# Beunruhigend große Fluktuationen beim Dornfelder?



# Beunruhigend große Fluktuationen beim Dornfelder?



# Rotwein in RLP: nur ein Tröpfchen?

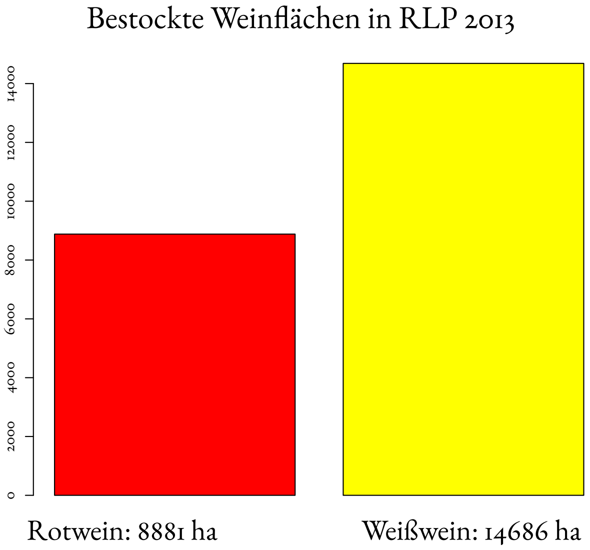
Bestockte Weinflächen in RLP 2013

Rotwein: 8881 ha

Weißwein: 14686 ha

Daten: Statistisches Landesamt RLP; Bilder (c) Benutzer André Karwath

# Rotwein in RLP: nur ein Tröpfchen?




Daten: Statistisches Landesamt RLP

# Fazit

- Histogramme erlauben einen detailierten Blick auf die Daten

# Fazit

- Histogramme erlauben einen detailierten Blick auf die Daten
- Dichtepolygone erlauben Vergleiche zwischen vielen Verteilungen

# Fazit

- Histogramme erlauben einen detailierten Blick auf die Daten
- Dichtepolygone erlauben Vergleiche zwischen vielen Verteilungen
- Boxplot können große Datenmengen vereinfacht zusammenfassen

# Fazit

- Histogramme erlauben einen detailierten Blick auf die Daten
- Dichtepolygone erlauben Vergleiche zwischen vielen Verteilungen
- Boxplot können große Datenmengen vereinfacht zusammenfassen
- Bei kleinen Datenmengen eher Stripcharts angemessen

# Fazit

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

# Inhalt

Es ist oft möglich,
das Wesentliche
an einer Stichprobe

mit ein paar Zahlen
zusammenzufassen.

# Wesentlich:

1. Wie groß?

2. Wie variabel?

# Wesentlich:

1. Wie groß?

Lageparameter

2. Wie variabel?

# Wesentlich:

1. Wie groß?

Lageparameter

2. Wie variabel?

Streuungsparameter

# Inhalt

# Eine Möglichkeit kennen wir schon aus dem Boxplot:

# Lageparameter

## Der Median

# Lageparameter

# Der Median

# Streuungsparameter

# Lageparameter

## Der Median

# Streuungsparameter

## Der Quartilabstand ($Q_3 - Q_1$)

# Der Median[1]:

# die Hälfte der Beobachtungen sind kleiner, die Hälfte sind größer.

[1] „saloppe" Definition (wir sehen gleich die präzise Definition)

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 Quartil[2], $Q_1$:

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

# Die Quartile

Das erste Quartil[2], $Q_1$:
ein Viertel der Beobachtungen
sind kleiner,
drei Viertel sind größer.

[2] „saloppe" Definition (wir sehen gleich die präzise Definition)

# Die Quartile

Das erste Quartil[2], $Q_1$:
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 Quartil[3], $Q_3$:

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

# Die Quartile

**Das dritte Quartil[3], $Q_3$:**
drei Viertel der Beobachtungen
sind kleiner,
ein Viertel sind größer.

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

# Die Quartile

Das dritte Quartil[3], $Q_3$:
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 $x_1, x_2, \ldots, x_n$ gegeben, $\alpha \in (0, 1)$.
$q$ ist (ein) $\alpha$-Quantil der $n$ Beobachtungswerte, wenn gilt

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

# (Empirische) Quantile, allgemein

Seien $n$ (reelle) Beobachtungswerte $x_1, x_2, \dots, x_n$ gegeben, $\alpha \in (0, 1)$.
$q$ ist (ein) $\alpha$-Quantil der $n$ Beobachtungswerte, wenn gilt

$$\frac{1}{n} \left|\{1 \le i \le n : x_i \le q\}\right| \ge \alpha \text{ und } \frac{1}{n} \left|\{1 \le i \le n : x_i \ge q\}\right| \ge 1 - \alpha.$$

**Bem.**: Im Allgemeinen ist ein $\alpha$-Quantil nicht eindeutig:

Seien $x_{(1)} \le x_{(2)} \le \cdots \le x_{(n)}$ die der Größe nach sortierten Werte.
Wenn $\alpha = \frac{k}{n}$ mit $1 \le k < n$, so ist jeder Wert $q \in [x_{(k)}, x_{(k+1)}]$ ein $\alpha$-Quantil, denn $\left|\{i : x_i \le x_{(k)}\}\right| \ge k$, $\left|\{i : x_i \ge x_{(k)}\}\right| \ge n - k + 1$.

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

# (Empirische) Quantile, allgemein II

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

(ein) $\alpha$-Quantil $q$ der Beobachtungswerte erfüllt

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

# (Empirische) Quantile, allgemein II

$n$ (reelle) Beobachtungswerte $x_1, x_2, \ldots, x_n$ gegeben, $\alpha \in (0,1)$.
(ein) $\alpha$-Quantil $q$ der Beobachtungswerte erfüllt

$$\frac{1}{n}\big|\{1 \le i \le n : x_i \le q\}\big| \ge \alpha \text{ und } \frac{1}{n}\big|\{1 \le i \le n : x_i \ge q\}\big| \ge 1-\alpha.$$

**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, vgl. Bem 1.13, 2. und Def. 3.22.
- In der Literatur (und auch in Statistik-Software) sind verschiedene Interpolationen üblich, um „das“ $\alpha$-Quantil stetig in $\alpha$ zu machen. (In `R` siehe etwa `help(quantile)`, es sind 9 Varianten implementiert.)
- Die Uneindeutigkeit des $\alpha$-Quantils ist für halbwegs große $n$ in der Praxis oft wenig von Belang.

# Inhalt

$n$ (reelle) Beobachtungswerte $x_1, x_2, \ldots, x_n$

Am häufigsten werden benutzt:

Lageparameter

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

$n$ (reelle) Beobachtungswerte $x_1, x_2, \ldots, x_n$

Am häufigsten werden benutzt:

Lageparameter

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

Streuungsparameter

Die Standardabweichung $s$ (bzw. $\sigma$)

$n$ (reelle) Beobachtungswerte $x_1, x_2, \ldots, x_n$

Am häufigsten werden benutzt:

Lageparameter

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

Streuungsparameter

Die Standardabweichung $s$ (bzw. $\sigma$)

wobei

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

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

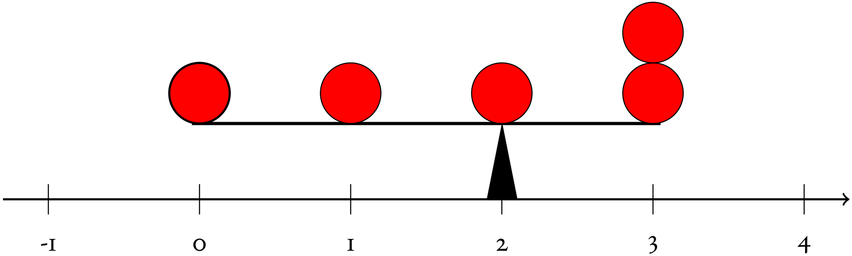
$\left( = \frac{n}{n-1} \sigma^2 \right)$

# Erinnerung: Geometrische Interpretation des Mittelwerts

# als Schwerpunkt

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

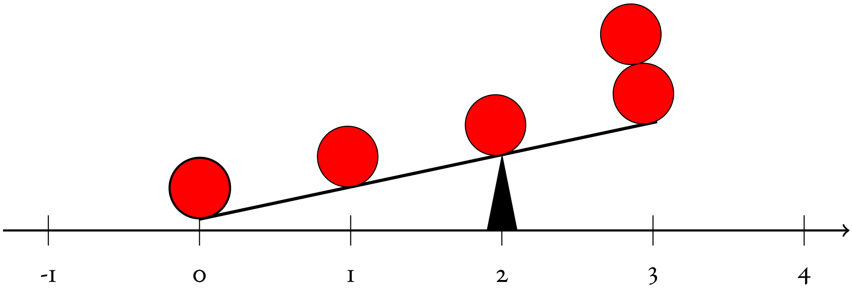
$$\bar{x} = 2\ ?$$




[3]Vgl. auch Bem. 3.2, 2.

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

$\overline{x} = 2$ ?

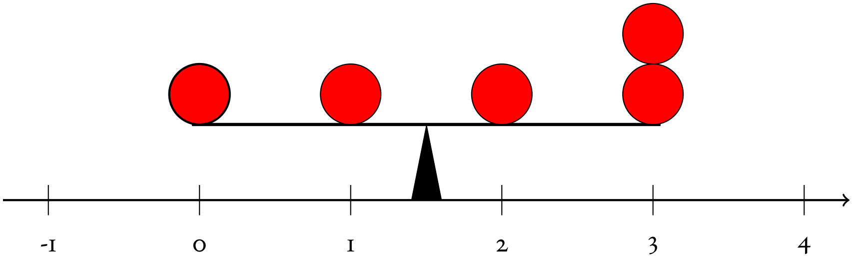



zu groß!

[3] Vgl. auch Bem. 3.2, 2.

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

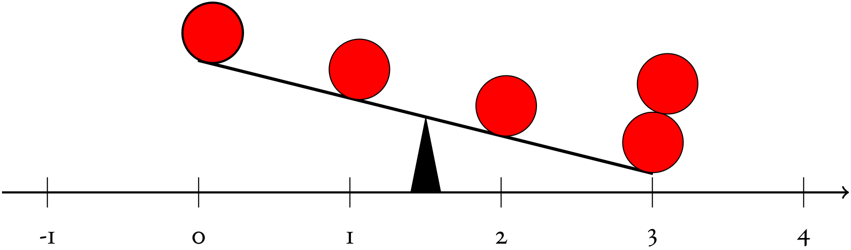
$$\overline{x} = 1.5 \,?$$




[3]Vgl. auch Bem. 3.2, 2.

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

$\overline{x} = 1.5$ ?

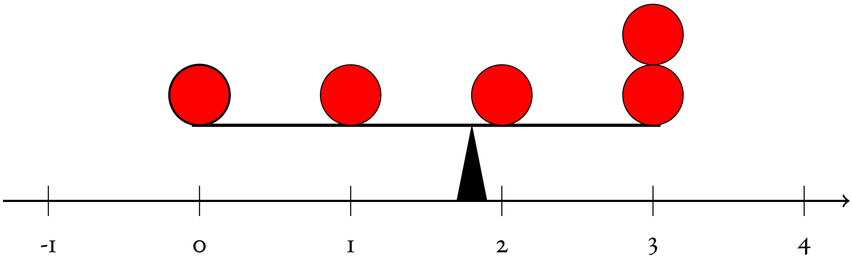



zu klein!

[3] Vgl. auch Bem. 3.2, 2.

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

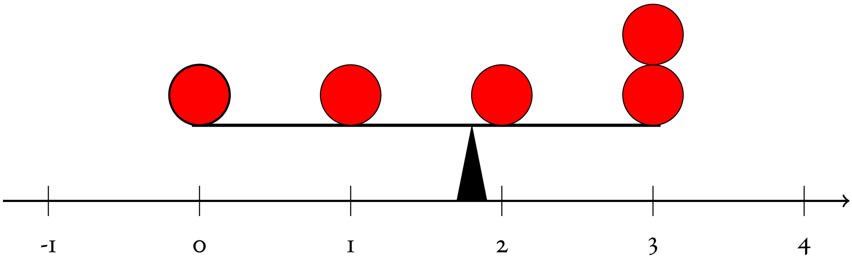
$$\overline{x} = 1.8\ ?$$




[3] Vgl. auch Bem. 3.2, 2.

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

$\overline{x} = 1.8$ ?




richtig!

[3]Vgl. auch Bem. 3.2, 2.

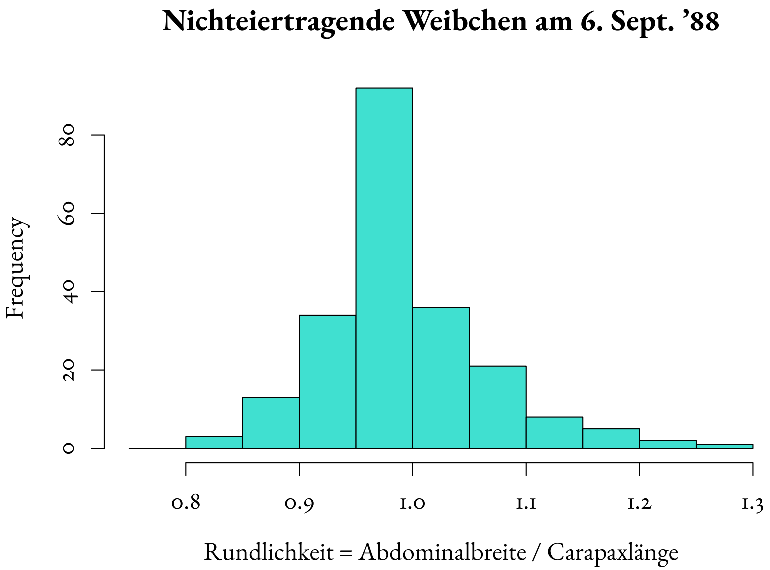
Oft kann man „mit dem bloßen Auge“ anhand eines Histogramms den Mittelwert gut einschätzen.

Beispiel: *Galathea intermedia*

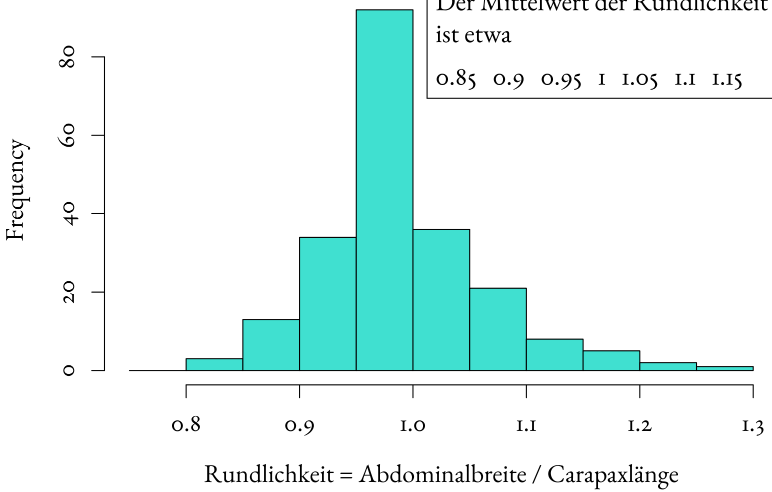
„Rundlichkeit“

:=

Abdominalbreite / Carapaxlänge


Nichteiertragende Weibchen am 6. Sept. ’88
Frequency
0
20
40
60
80
0.8
0.9
1.0
1.1
1.2
1.3
Rundlichkeit = Abdominalbreite / Carapaxlänge

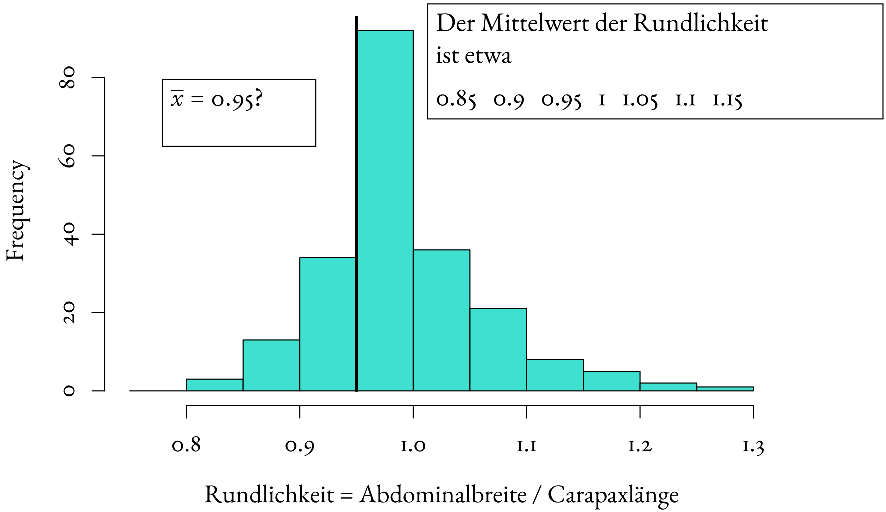
## Nichteiertragende Weibchen am 6. Sept. ’88

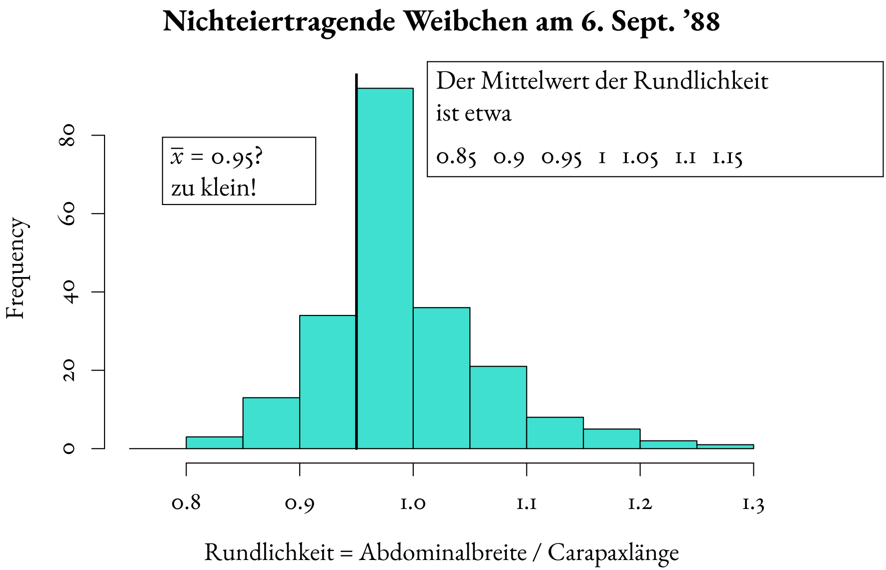



Der Mittelwert der Rundlichkeit ist etwa

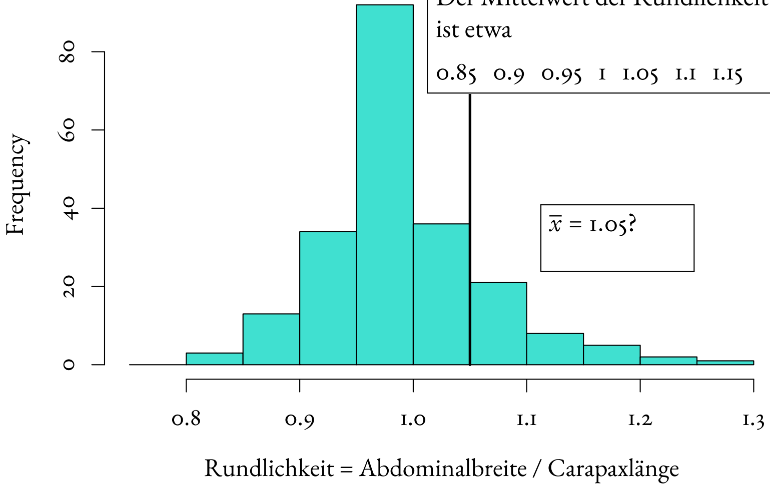
0.85   0.9   0.95   1   1.05   1.1   1.15

## Nichteiertragende Weibchen am 6. Sept. ’88




Nichteiertragende Weibchen am 6. Sept. ’88
$\overline{x} = 0.95$?
zu klein!
Der Mittelwert der Rundlichkeit
ist etwa
0.85 0.9 0.95 1 1.05 1.1 1.15
Frequency
0
20
40
60
80
0.8
0.9
1.0
1.1
1.2
1.3
Rundlichkeit = Abdominalbreite / Carapaxlänge

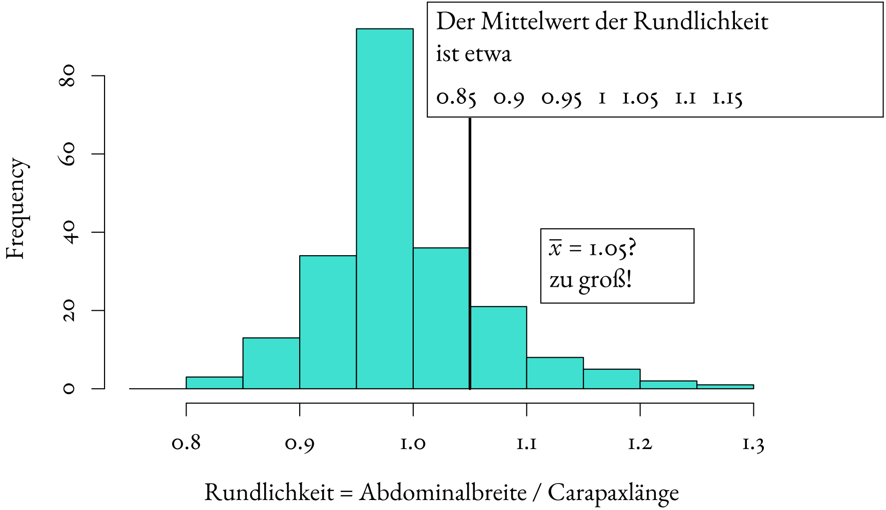
## Nichteiertragende Weibchen am 6. Sept. ’88



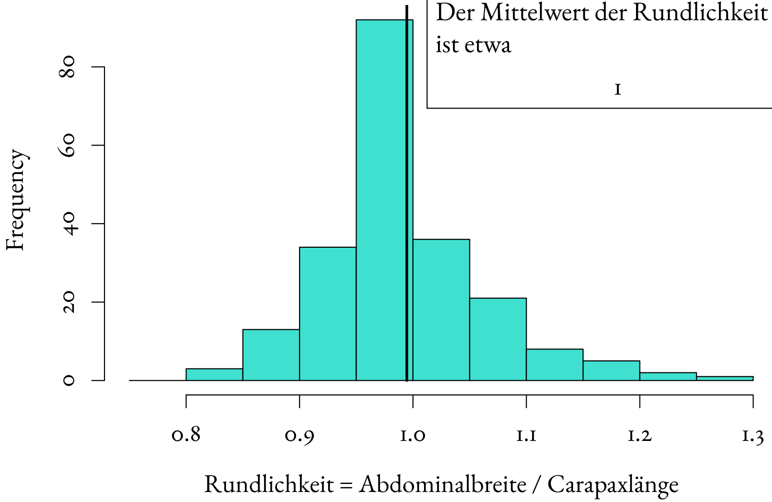

Der Mittelwert der Rundlichkeit ist etwa

0.85 0.9 0.95 1 1.05 1.1 1.15

## Nichteiertragende Weibchen am 6. Sept. ’88

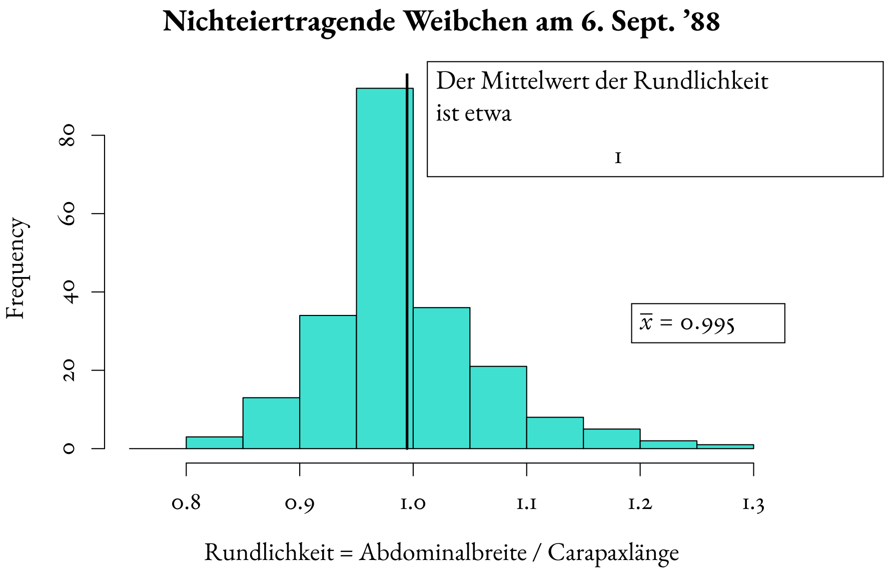


## Nichteiertragende Weibchen am 6. Sept. ’88



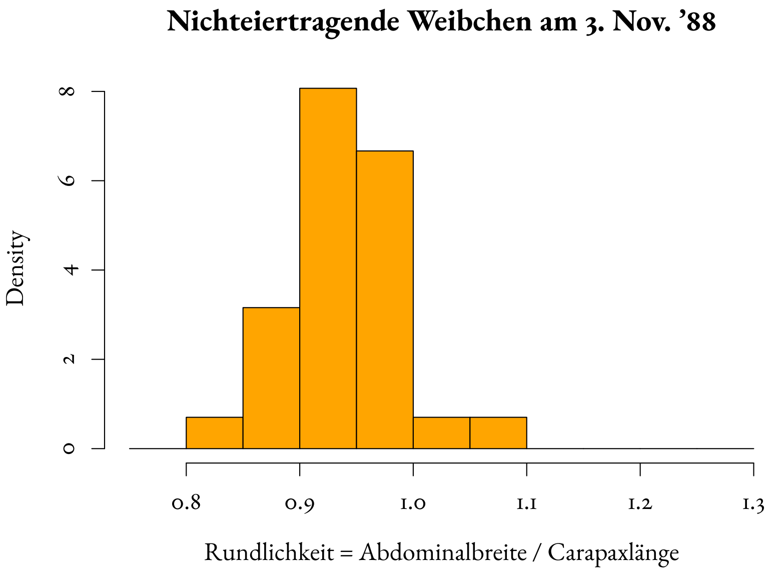

Der Mittelwert der Rundlichkeit ist etwa

1

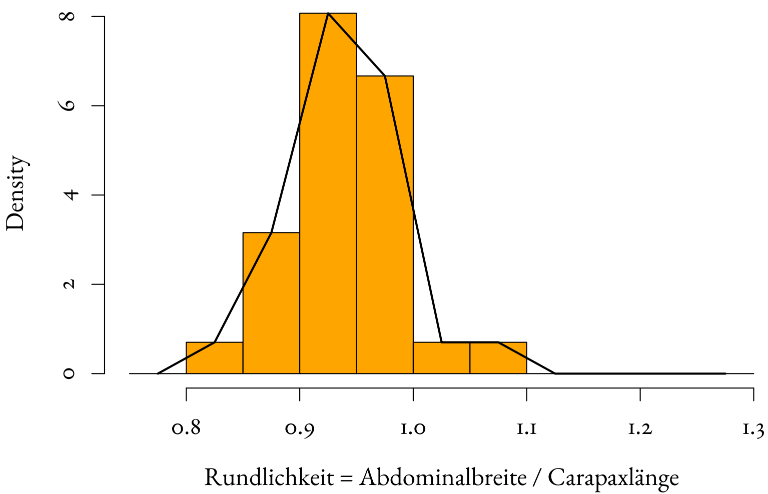

Nichteiertragende Weibchen am 6. Sept. ’88
Der Mittelwert der Rundlichkeit
ist etwa
1
$\overline{x}$ = 0.995
Frequency
0
20
40
60
80
0.8
0.9
1.0
1.1
1.2
1.3
Rundlichkeit = Abdominalbreite / Carapaxlänge

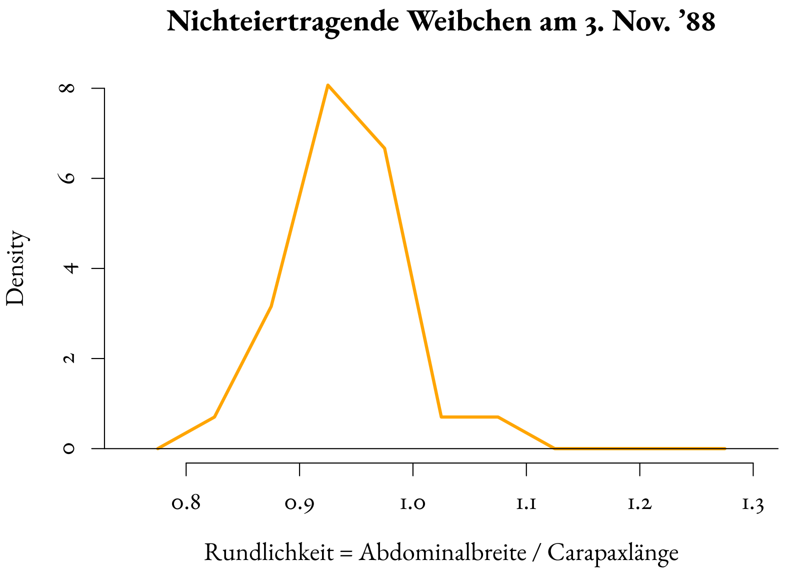
# Noch ein Beispiel:

## Die Daten vom 3.11.88

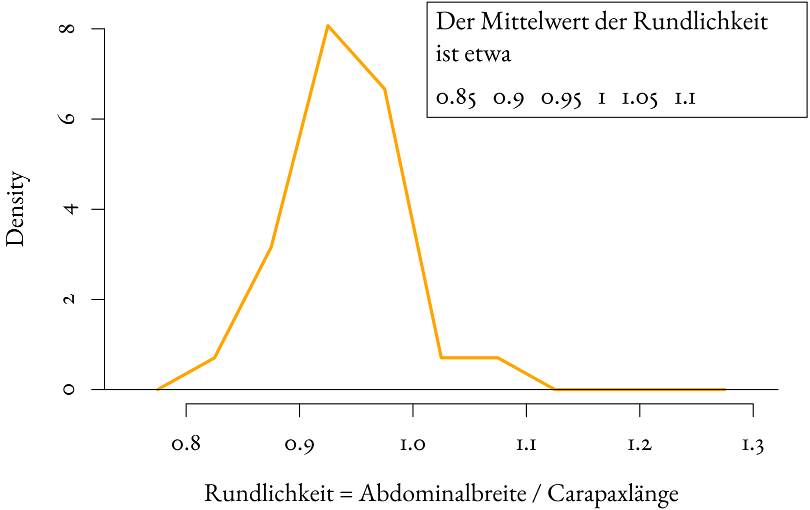

Nichteiertragende Weibchen am 3. Nov. '88
Density
0
2
4
6
8
0.8
0.9
1.0
1.1
1.2
1.3
Rundlichkeit = Abdominalbreite / Carapaxlänge

## Nichteiertragende Weibchen am 3. Nov. '88

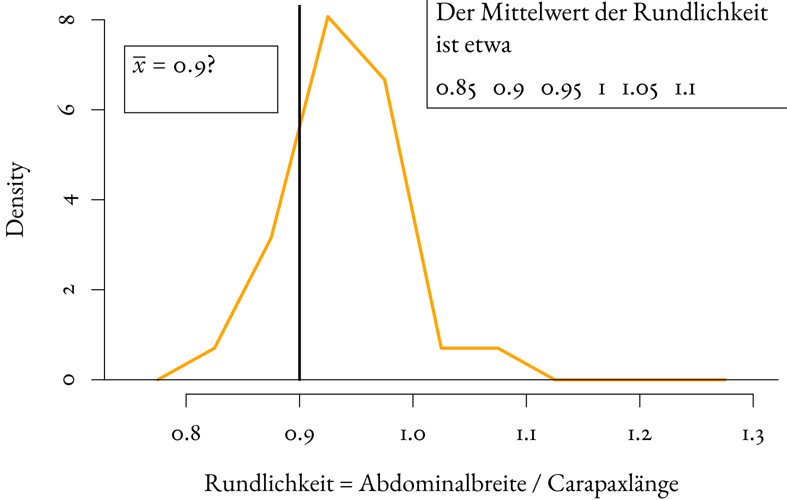



Nichteiertragende Weibchen am 3. Nov. ’88
Density
0
2
4
6
8
0.8
0.9
1.0
1.1
1.2
1.3
Rundlichkeit = Abdominalbreite / Carapaxlänge

# Nichteiertragende Weibchen am 3. Nov. '88



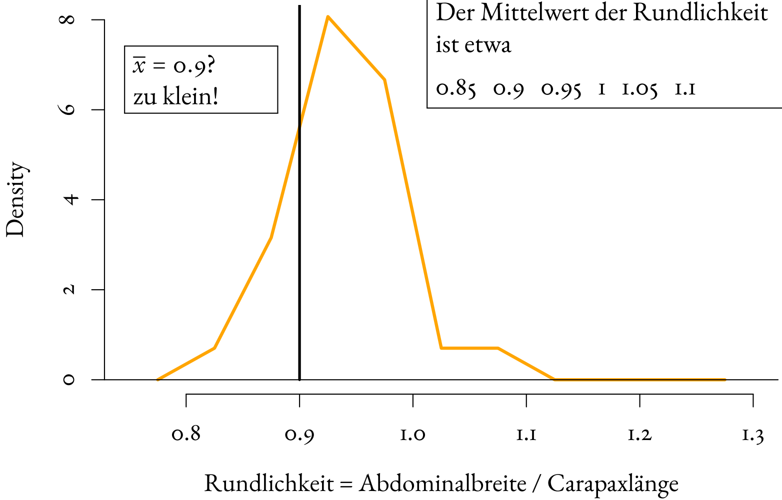
## Nichteiertragende Weibchen am 3. Nov. '88




Der Mittelwert der Rundlichkeit ist etwa

0.85 0.9 0.95 1 1.05 1.1

# Nichteiertragende Weibchen am 3. Nov. '88

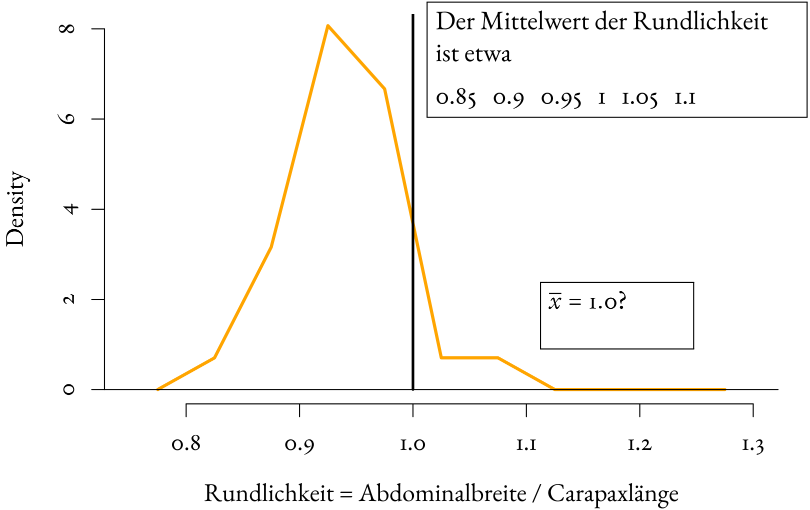


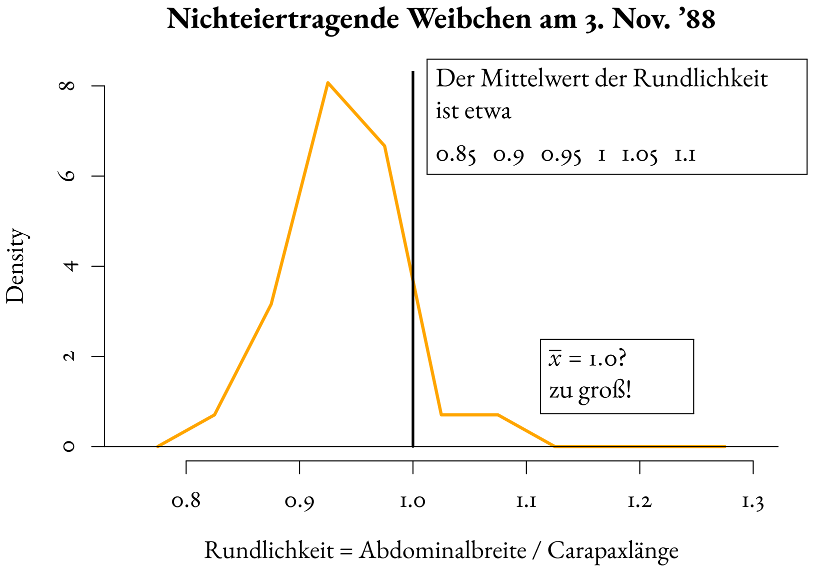

$\bar{x} = 0.9$?
zu klein!

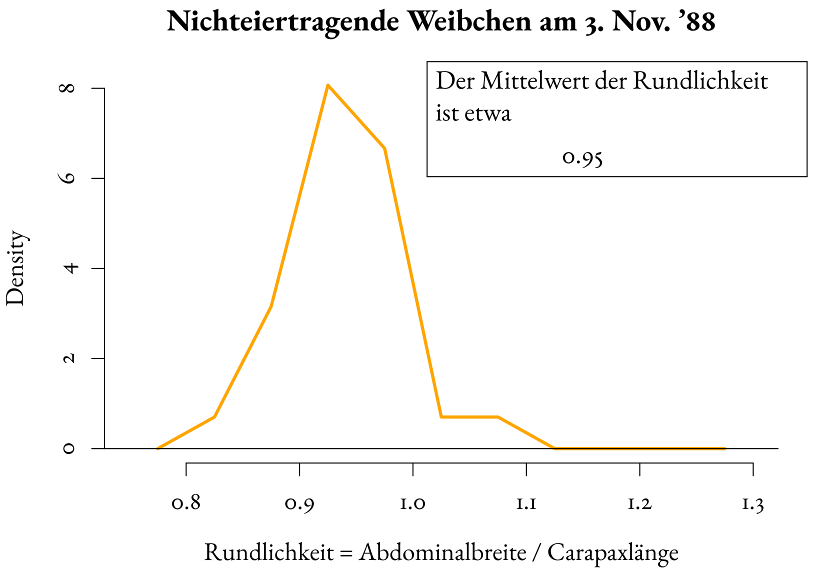
Der Mittelwert der Rundlichkeit ist etwa

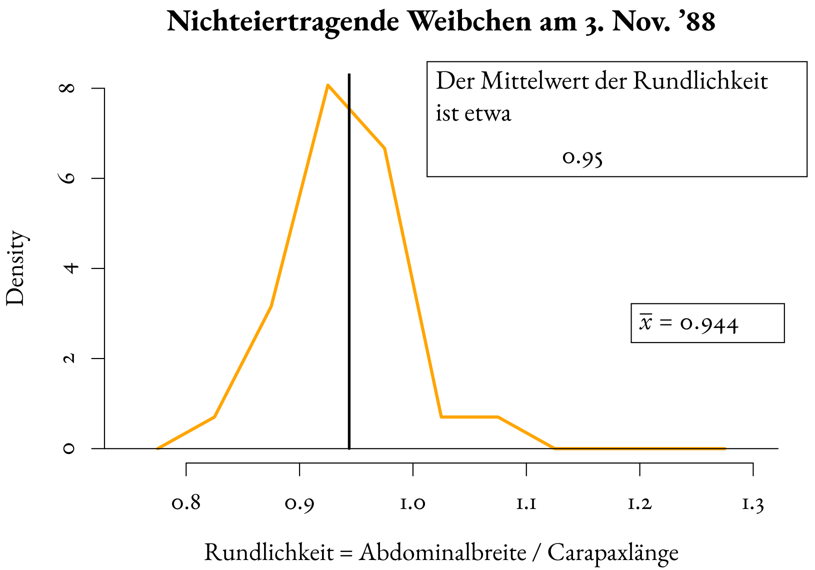
0.85 0.9 0.95 1 1.05 1.1

## Nichteiertragende Weibchen am 3. Nov. ’88




Nichteiertragende Weibchen am 3. Nov. ’88
Der Mittelwert der Rundlichkeit ist etwa
0.85 0.9 0.95 1 1.05 1.1
$\overline{x}$ = 1.0?
zu groß!
Density
0
2
4
6
8
0.8
0.9
1.0
1.1
1.2
1.3
Rundlichkeit = Abdominalbreite / Carapaxlänge


Nichteiertragende Weibchen am 3. Nov. ’88
Der Mittelwert der Rundlichkeit ist etwa
0.95
Density
0
2
4
6
8
0.8
0.9
1.0
1.1
1.2
1.3
Rundlichkeit = Abdominalbreite / Carapaxlänge


Nichteiertragende Weibchen am 3. Nov. '88
Der Mittelwert der Rundlichkeit ist etwa
0.95
$\overline{x} = 0.944$
Density
0
2
4
6
8
0.8
0.9
1.0
1.1
1.2
1.3
Rundlichkeit = Abdominalbreite / Carapaxlänge

# Die Standardabweichung (auch: Streuung)

# Die Standardabweichung (auch: Streuung)

Wie weit weicht
eine typische Beobachtung
vom
Mittelwert
ab ?

# Mit $n$ oder $n-1$ rechnen?

Die Standardabweichung $\sigma$ einer Zufallsvariable mit $n$ gleichwahrscheinlichen Ausgängen $x_1, \ldots, x_n$ (z.B. Würfelwurf) ist definiert durch (vgl. Def. 3.14)

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

# Mit $n$ oder $n-1$ rechnen?

Die Standardabweichung $\sigma$ einer Zufallsvariable mit $n$ gleichwahrscheinlichen Ausgängen $x_1, \ldots, x_n$ (z.B. Würfelwurf) ist definiert durch (vgl. Def. 3.14)

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

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

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

# $s$ als Schätzer für $\sigma$

Wir werden sehen:

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

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

so hat die *Zufallsvariable*

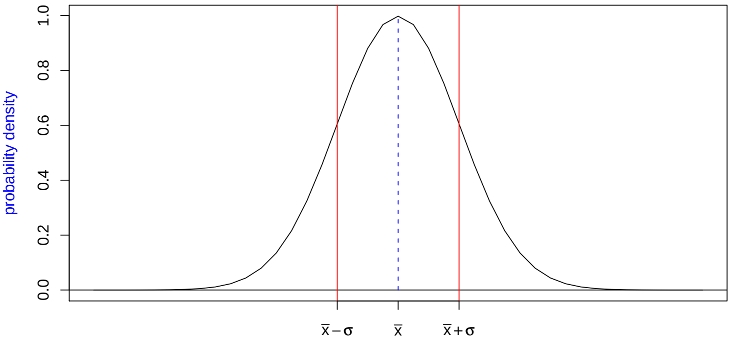
$$S^2 := \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \overline{X})^2$$

die Eigenschaft

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

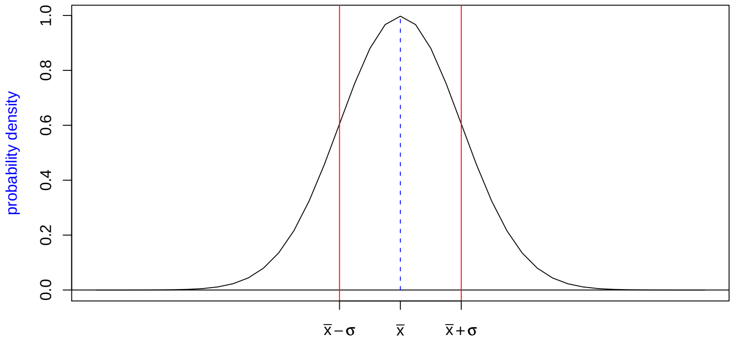
# Faustregel für die Standardabweichung

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



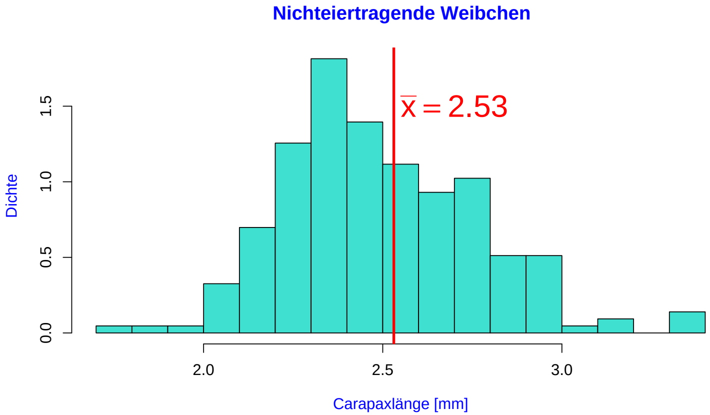
# Faustregel für die Standardabweichung

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

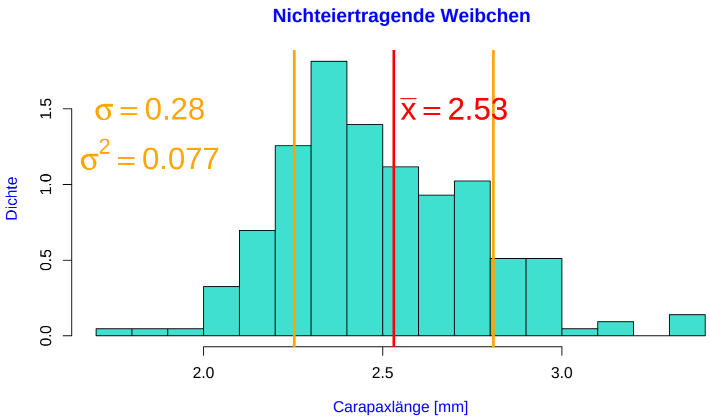



Oft kann man so die Standardabweichung „mit bloßem Auge“ abschätzen.

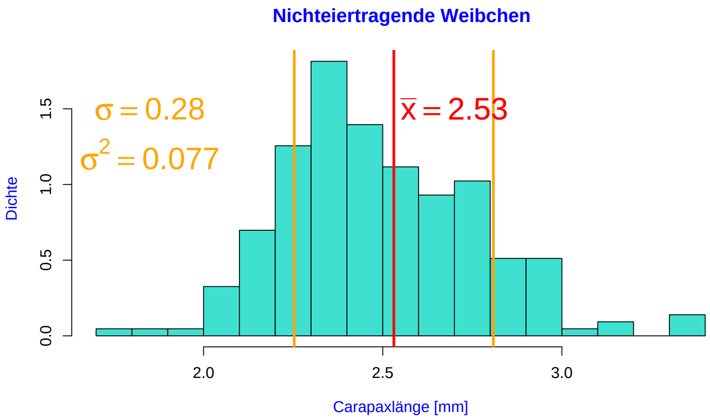
# Standardabweichung der Carapaxlängen nichteiertragender Weibchen vom 6.9.88



# Standardabweichung der Carapaxlängen nichteiertragender Weibchen vom 6.9.88



# Standardabweichung der Carapaxlängen nichteiertragender Weibchen vom 6.9.88




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

# Inhalt

Mittelwert und Standardabweichung. . .

- charakterisieren die Daten gut, falls deren Verteilung (zumindest in etwa) glockenförmig ist

Mittelwert und Standardabweichung. . .

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

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.)

# Inhalt

# Bachstelzen fressen Dungfliegen

Räuber

Bachstelze (White Wagtail)
*Motacilla alba alba*

image (c) by Artur Mikołajewski

Beute

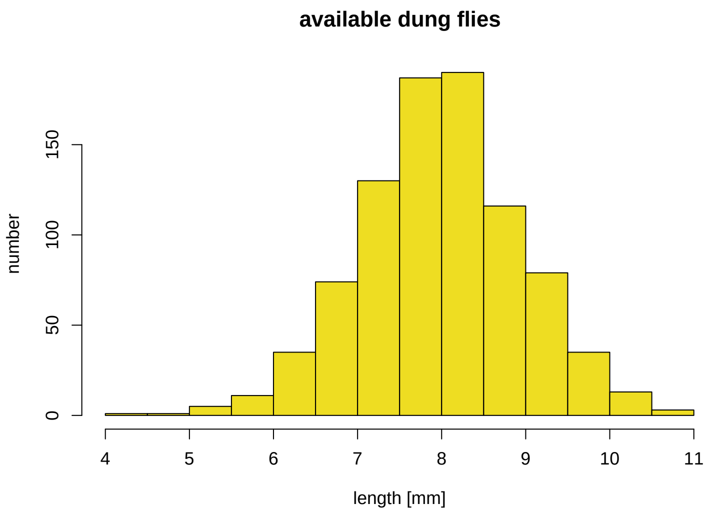
Gelbe Dungfliege
*Scatophaga stercoraria*

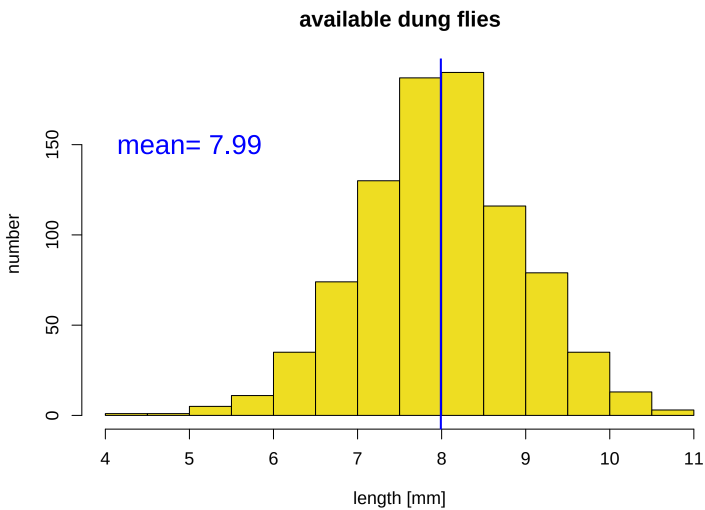
image (c) by Viatour Luc

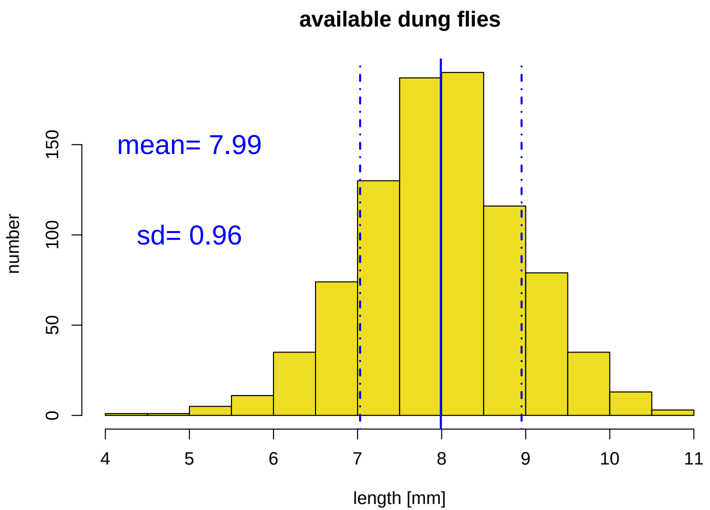
# Vermutung

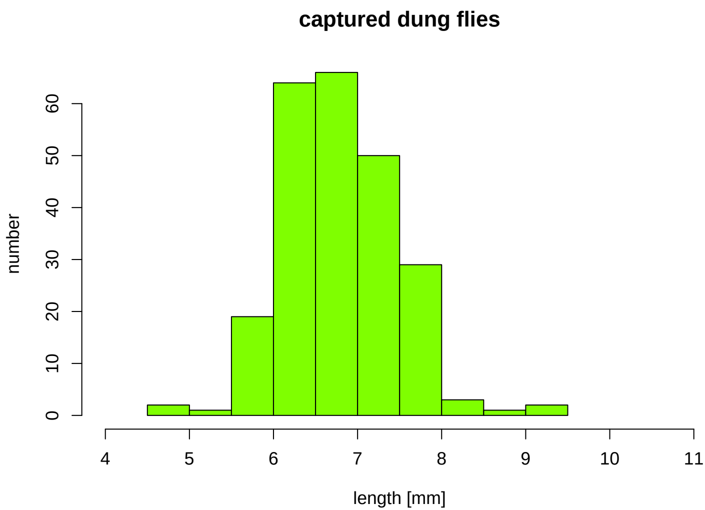
- Die Fliegen sind unterschiedlich groß
- Effizienz für die Bachstelze = Energiegewinn / Zeit zum Fangen und fressen
- Laborexperimente lassen vermuten, dass die Effizienz bei 7mm großen Fliegen maximal ist.

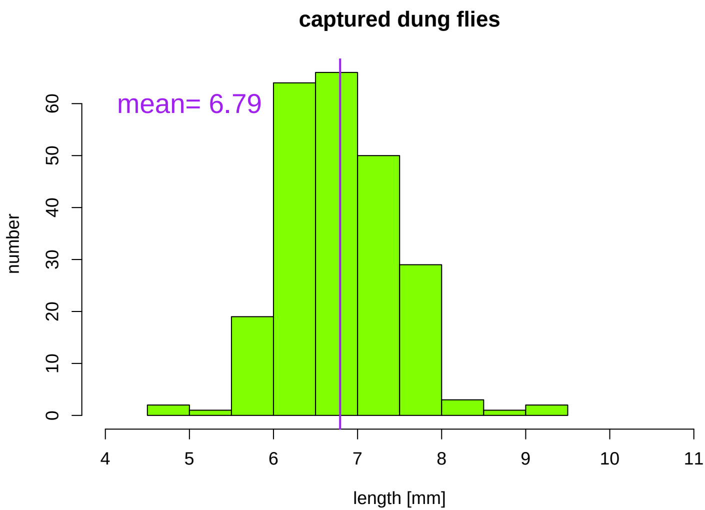
N.B. Davies.
Prey selection and social behaviour in wagtails (Aves: Motacillidae).
*J. Anim. Ecol.*, 46:37–57, 1977.

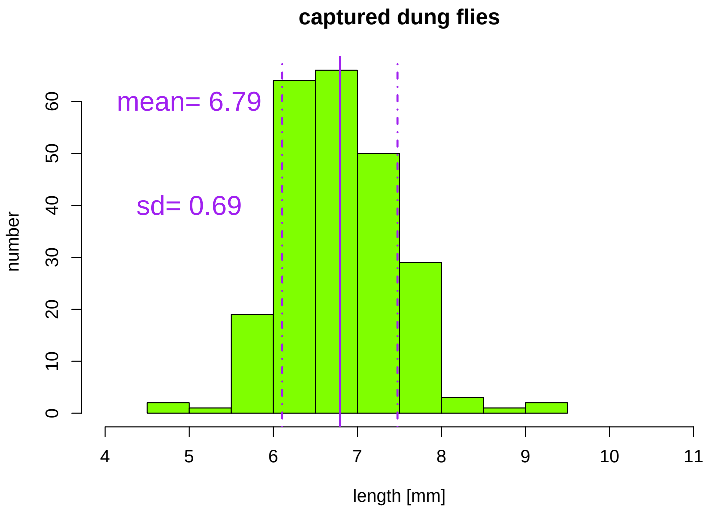

available dung flies
number
0
50
100
150
4
5
6
7
8
9
10
11
length [mm]

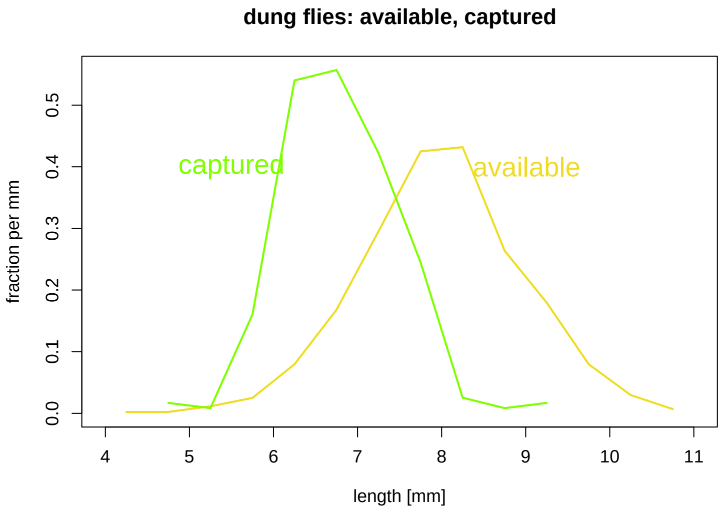

available dung flies
mean= 7.99
number
0
50
100
150
4
5
6
7
8
9
10
11
length [mm]


available dung flies
mean= 7.99
sd= 0.96
number
0
50
100
150
4
5
6
7
8
9
10
11
length [mm]


captured dung flies
number
0
10
20
30
40
50
60
4
5
6
7
8
9
10
11
length [mm]

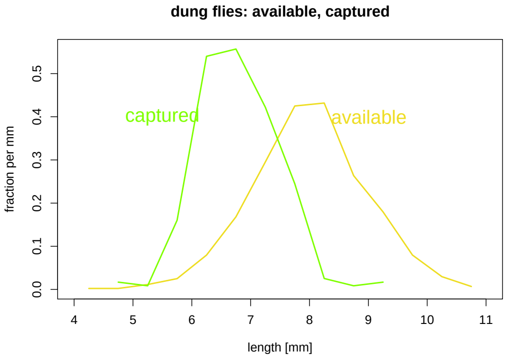

captured dung flies
mean= 6.79
number
0
10
20
30
40
50
60
4
5
6
7
8
9
10
11
length [mm]


captured dung flies
mean= 6.79
sd= 0.69
number
0
10
20
30
40
50
60
4
5
6
7
8
9
10
11
length [mm]


dung flies: available, captured
captured
available
fraction per mm
0.0
0.1
0.2
0.3
0.4
0.5
4
5
6
7
8
9
10
11
length [mm]

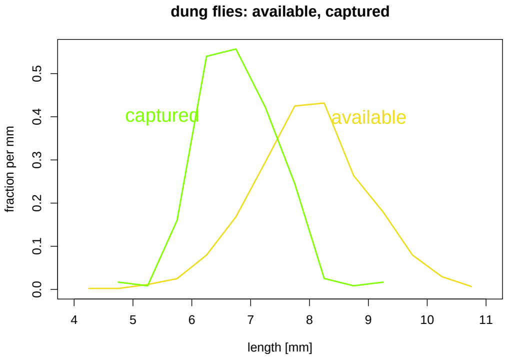
# Vergleich der Größenverteilungen

| | captured | available |
|---|---|---|
| Mittelwert | | |



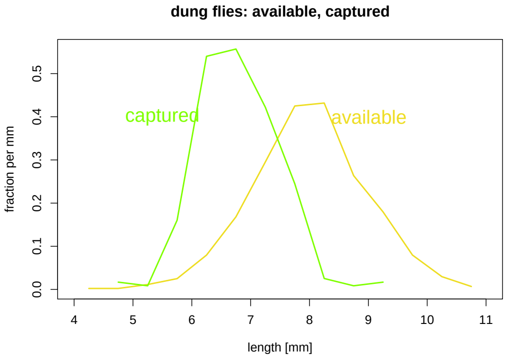
# Vergleich der Größenverteilungen

| | captured | available |
|---|---|---|
| Mittelwert | < | |



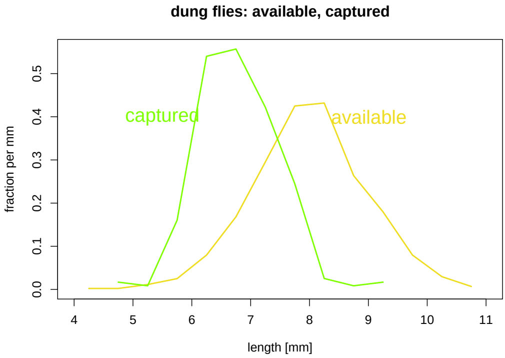
# Vergleich der Größenverteilungen

| | captured | | available |
|---|---|---|---|
| Mittelwert | 6.79 | < | 7.99 |



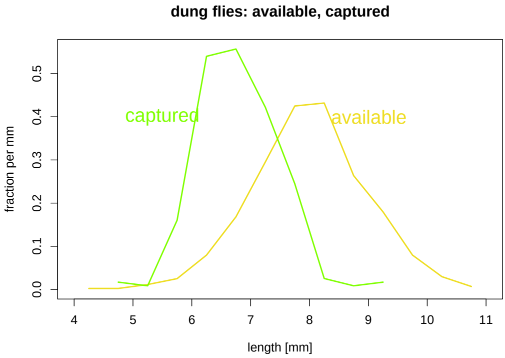
# Vergleich der Größenverteilungen

| | captured | | available |
|---|---|---|---|
| Mittelwert | 6.79 | < | 7.99 |
| Standardabweichung | | | |



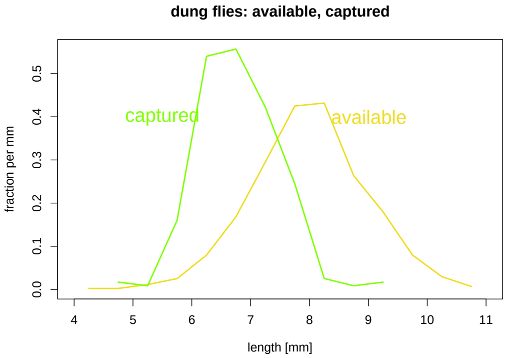
# Vergleich der Größenverteilungen

| | captured | | available |
|---|---|---|---|
| Mittelwert | 6.79 | < | 7.99 |
| Standardabweichung | | < | |



# Vergleich der Größenverteilungen

| | captured | | available |
|---|---|---|---|
| Mittelwert | 6.79 | < | 7.99 |
| Standardabweichung | 0.69 | < | 0.96 |



# Interpretation

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

# Interpretation

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

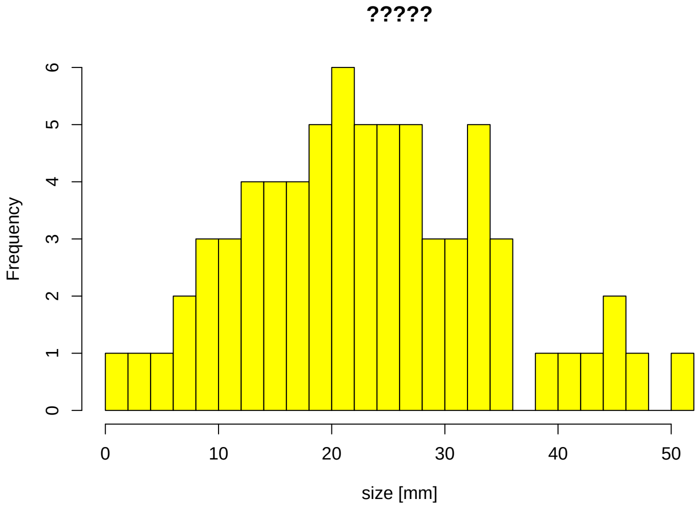
Hier waren die Verteilungen glockenförmig und es genügten 4 Werte (die beiden Mittelwerte und die beiden Standardabweichungen), um die Daten adäquat zu beschreiben.

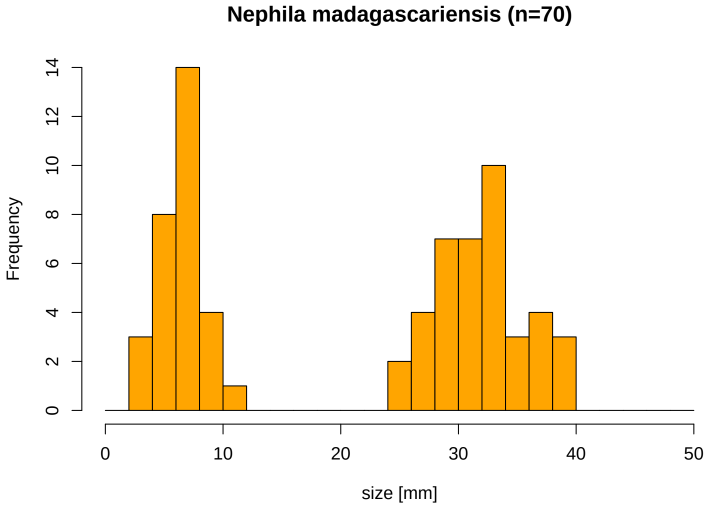
# Inhalt

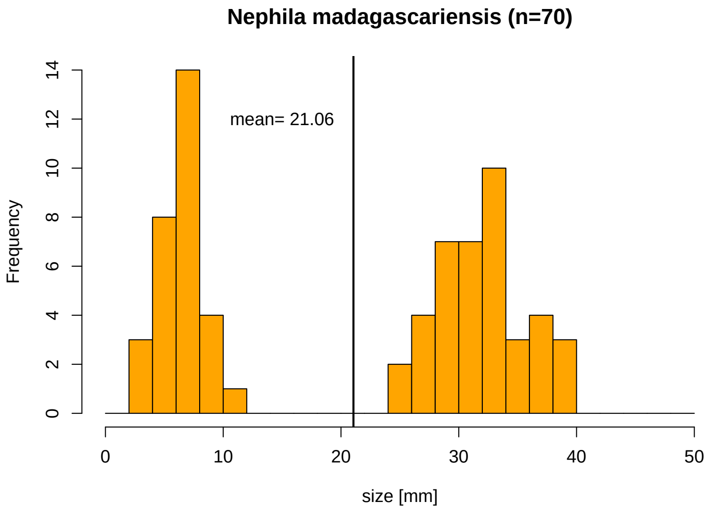
*Nephila madagascariensis*
image (c) by Bernard Gagnon

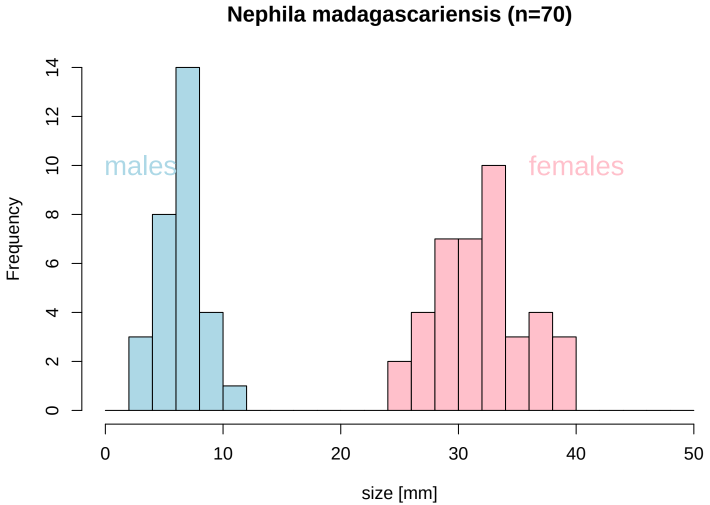
## Simulierte Daten:

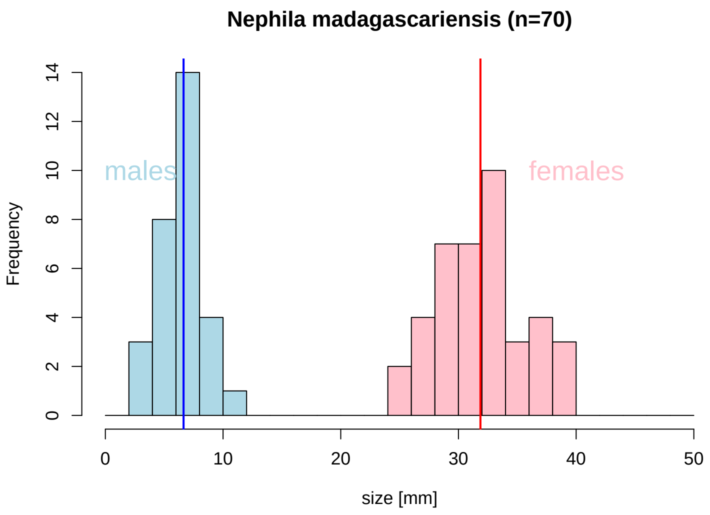
Eine Stichprobe von 70 Spinnen
Mittlere Größe: 21,06 mm
Standardabweichung der Größe: 12,94 mm


?????
Frequency
0
1
2
3
4
5
6
0
10
20
30
40
50
size [mm]


Nephila madagascariensis (n=70)
Frequency
0
2
4
6
8
10
12
14
0
10
20
30
40
50
size [mm]


Nephila madagascariensis (n=70)
mean= 21.06
Frequency
0
2
4
6
8
10
12
14
0
10
20
30
40
50
size [mm]


Nephila madagascariensis (n=70)
males
females
Frequency
0
2
4
6
8
10
12
14
0
10
20
30
40
50
size [mm]


Nephila madagascariensis (n=70)
males
females
Frequency
0
2
4
6
8
10
12
14
0
10
20
30
40
50
size [mm]

*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.

# Inhalt

# 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.

A.D. Bradshaw.
Population Differentiation in *agrostis tenius 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.
*Evolution*, 22:108–118, 1968.

# Anpassung an Kupfer?

- Pflanzen, denen das Kupfer schadet, haben kürzere Wurzeln.
- Die Wurzellängen von Pflanzen aus der Umgebung von Kupferminen wird gemessen.

A.D. Bradshaw.
Population Differentiation in *agrostis tenius 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.
*Evolution*, 22:108–118, 1968.

# 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.

A.D. Bradshaw.
Population Differentiation in *agrostis tenius Sibth*. III. populations in varied environments.
*New Phytologist*, 59(1):92 – 103, 1960.

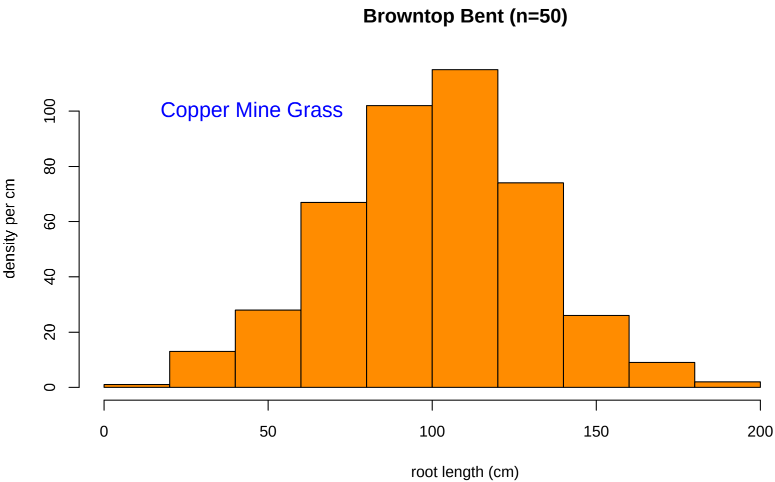
T. McNeilly and A.D Bradshaw.
Evolutionary Processes in Populations of Copper Tolerant Agrostis tenuis Sibth.
*Evolution*, 22:108–118, 1968.

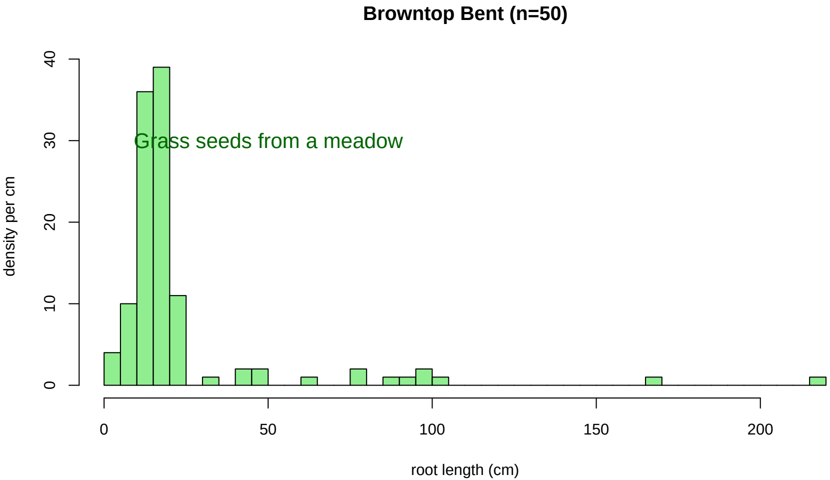
# Anpassung an Kupfer?

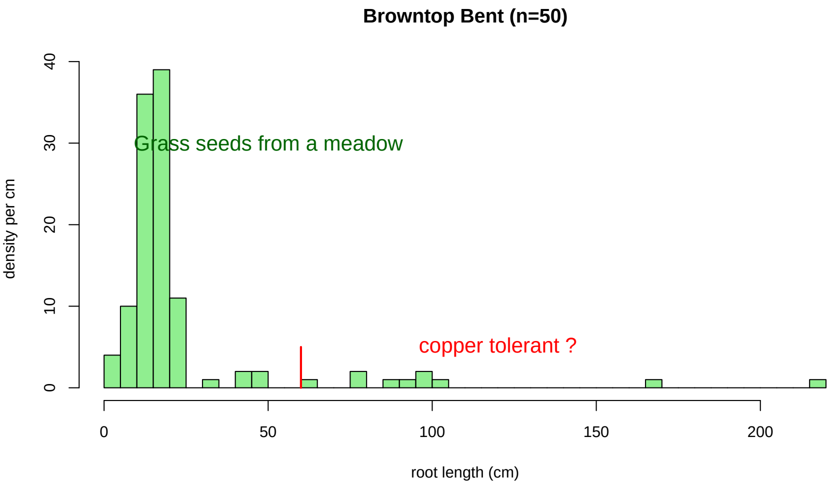
- Pflanzen, denen das Kupfer schadet, haben kürzere Wurzeln.
- Die Wurzellängen von Pflanzen aus der Umgebung von Kupferminen wird gemessen.
- Samen von unbelasteten Wiesen werden bei Kupferminen eingesät.
- Die Wurzellängen dieser "Wiesenpflanzen" werden gemessen.

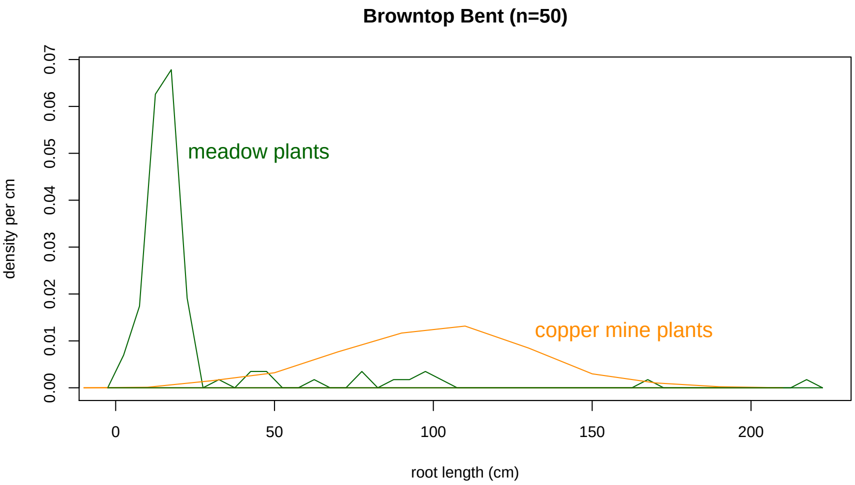
A.D. Bradshaw.
Population Differentiation in *agrostis tenius Sibth*. III. populations in varied environments.
*New Phytologist*, 59(1):92 – 103, 1960.

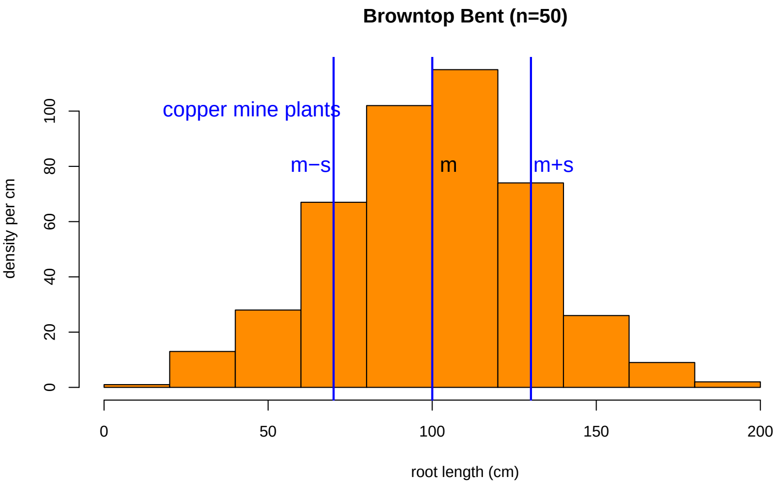
T. McNeilly and A.D Bradshaw.
Evolutionary Processes in Populations of Copper Tolerant Agrostis tenuis Sibth.
*Evolution*, 22:108–118, 1968.

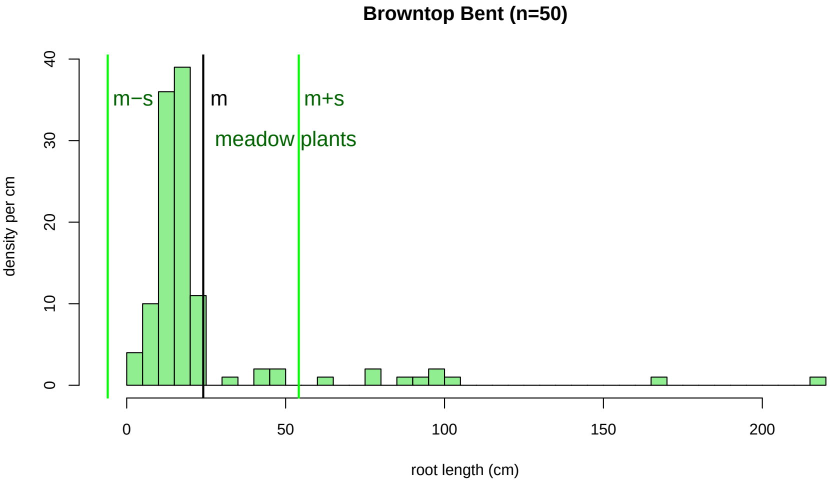

Browntop Bent (n=50)
Copper Mine Grass
density per cm
0
20
40
60
80
100
0
50
100
150
200
root length (cm)


Browntop Bent (n=50)
Grass seeds from a meadow
density per cm
0
10
20
30
40
0
50
100
150
200
root length (cm)


Browntop Bent (n=50)
Grass seeds from a meadow
copper tolerant ?
density per cm
0
10
20
30
40
0
50
100
150
200
root length (cm)


Browntop Bent (n=50)
meadow plants
copper mine plants
density per cm
0.00
0.01
0.02
0.03
0.04
0.05
0.06
0.07
0
50
100
150
200
root length (cm)


Browntop Bent (n=50)
copper mine plants
m−s
m
m+s
density per cm
0
20
40
60
80
100
0
50
100
150
200
root length (cm)




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

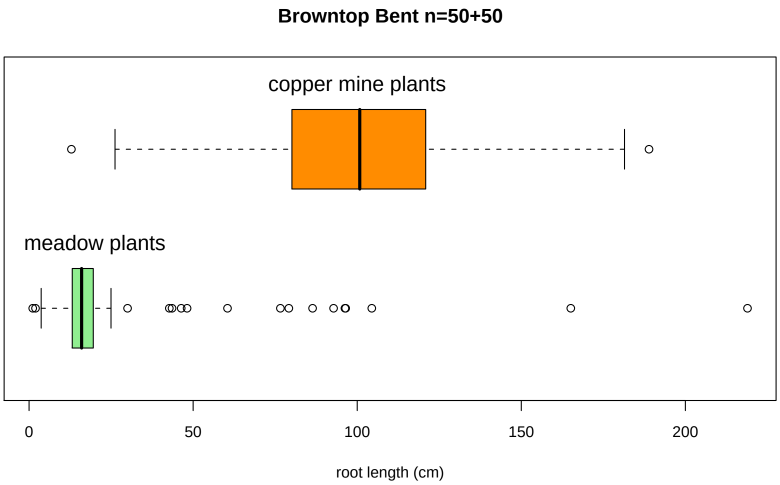
# Fazit des Straußgras-Beispiels

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

# 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
copper mine plants
meadow plants
0
50
100
150
200
root length (cm)

# Schlussfolgerung

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

# 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.