7. Ideen aus der Statistik

7.1 Deskriptive Statistik

Matthias Birkner

http://www.mathematik.uni-mainz.de/~birkner/GrundlStoch_1314/

13.1.2014

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 Natur ist voller Variabilität.

Worum geht es in der Statistik?

Die Natur ist voller Variabilität.

Wie geht man mit variablen Daten um?

Worum geht es in der Statistik?

Die Natur 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 Natur 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 Natur 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

Der Springkrebs *Galathea intermedia*

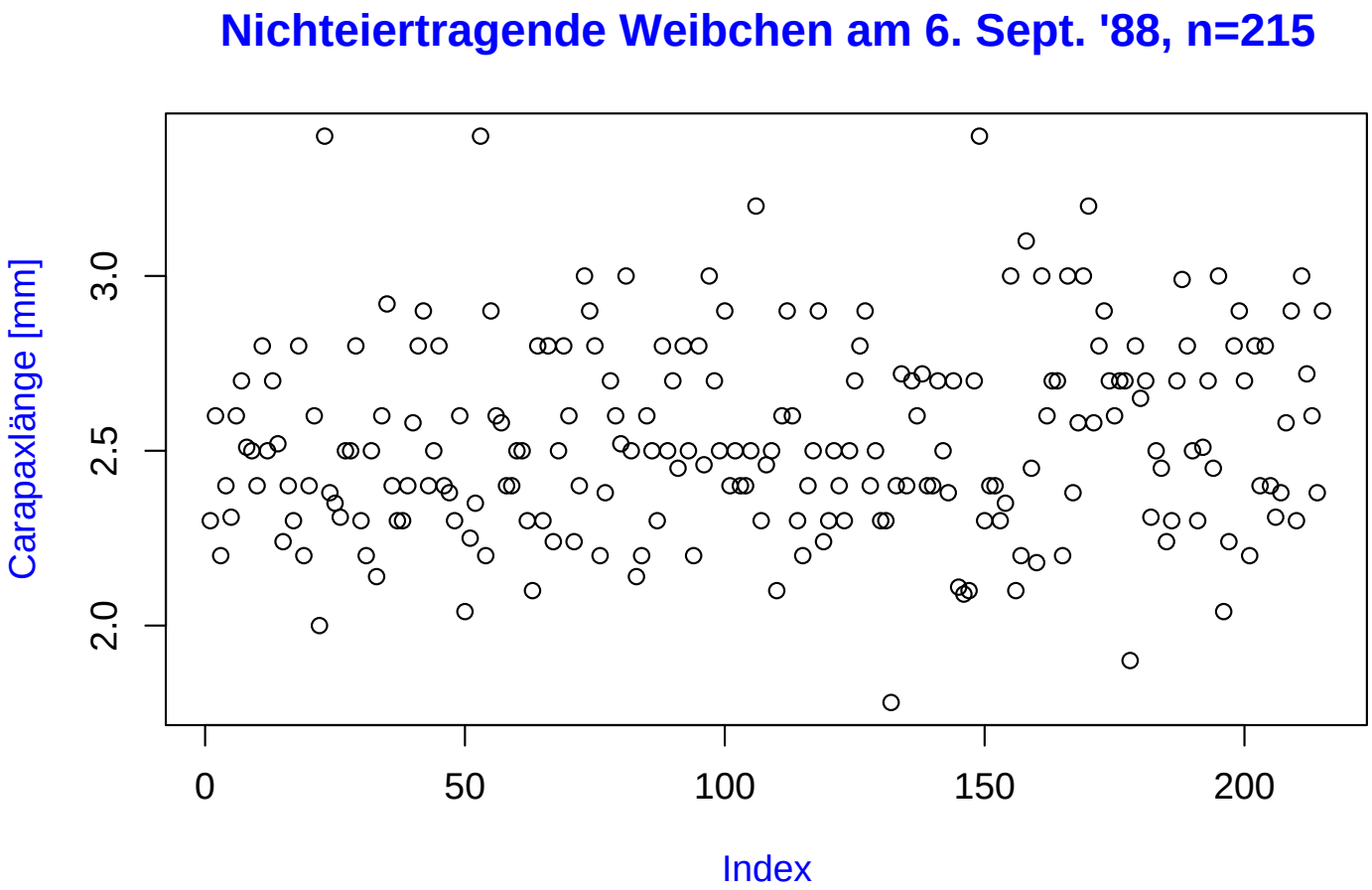
(Daten aus einer Diplomarbeit aus 2001 am Forschungsinstitut Senckenberg, Frankfurt am Main, Crustaceensektion, Leitung: Prof. Michael Türkay)

Helgoländer Tiefe Rinne, Fang vom 6.9.1988

Carapaxlänge (mm):

Nichteiertragende Weibchen ($n = 215$)

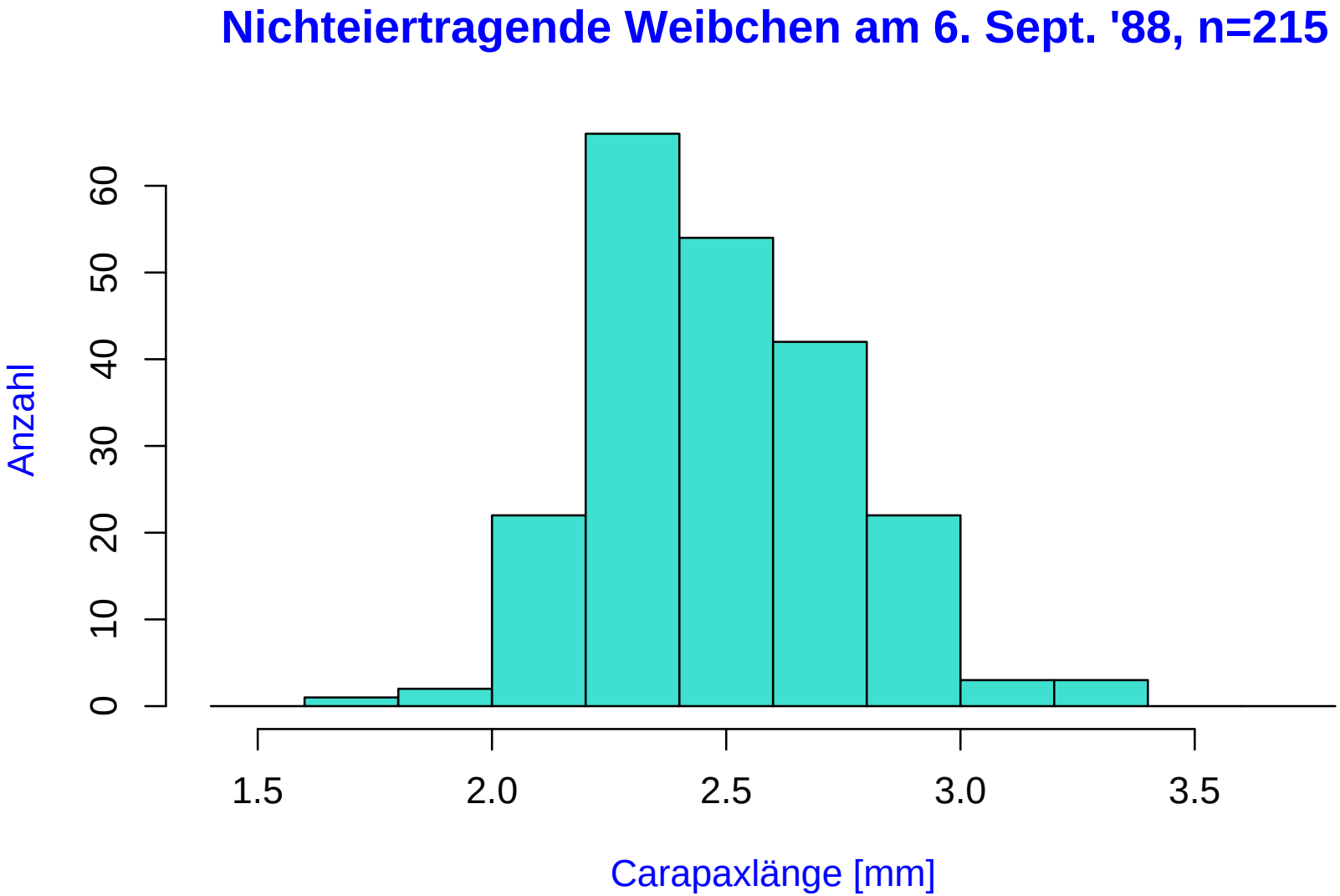
2,9	3,0	2,9	2,5	2,7	2,9	2,9	3,0
3,0	2,9	3,4	2,8	2,9	2,8	2,8	2,4
2,8	2,5	2,7	3,0	2,9	3,2	3,1	3,0
2,7	2,5	3,0	2,8	2,8	2,8	2,7	3,0
2,6	3,0	2,9	2,8	2,9	2,9	2,3	2,7
2,6	2,7	2,5	.	.	.	.	.


Nichteiertragende Weibchen am 6. Sept. '88, n=215
Carapaxlänge [mm]
2.0
2.5
3.0
0
50
100
150
200
Index

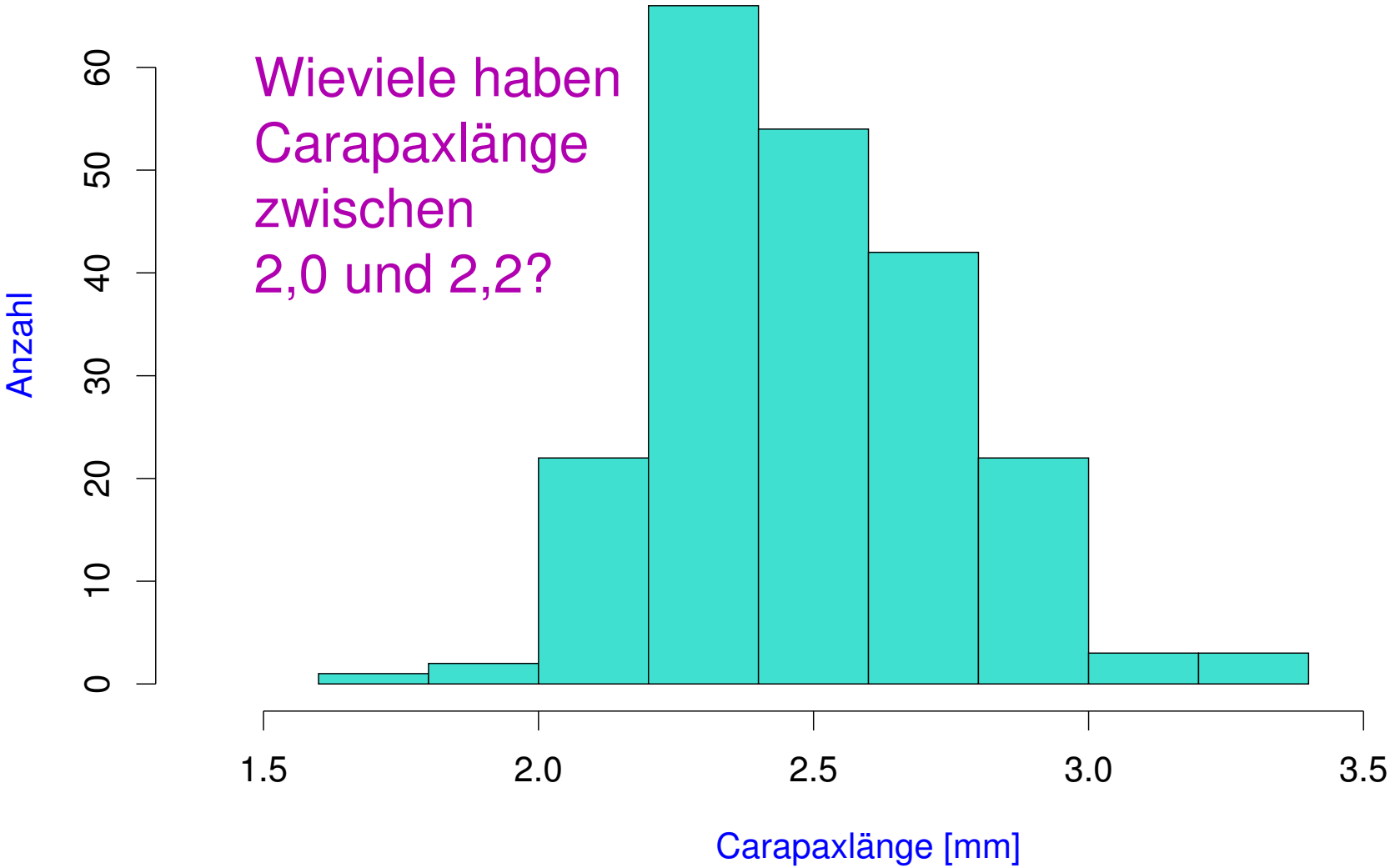
Inhalt

Eine Möglichkeit der graphischen Darstellung:

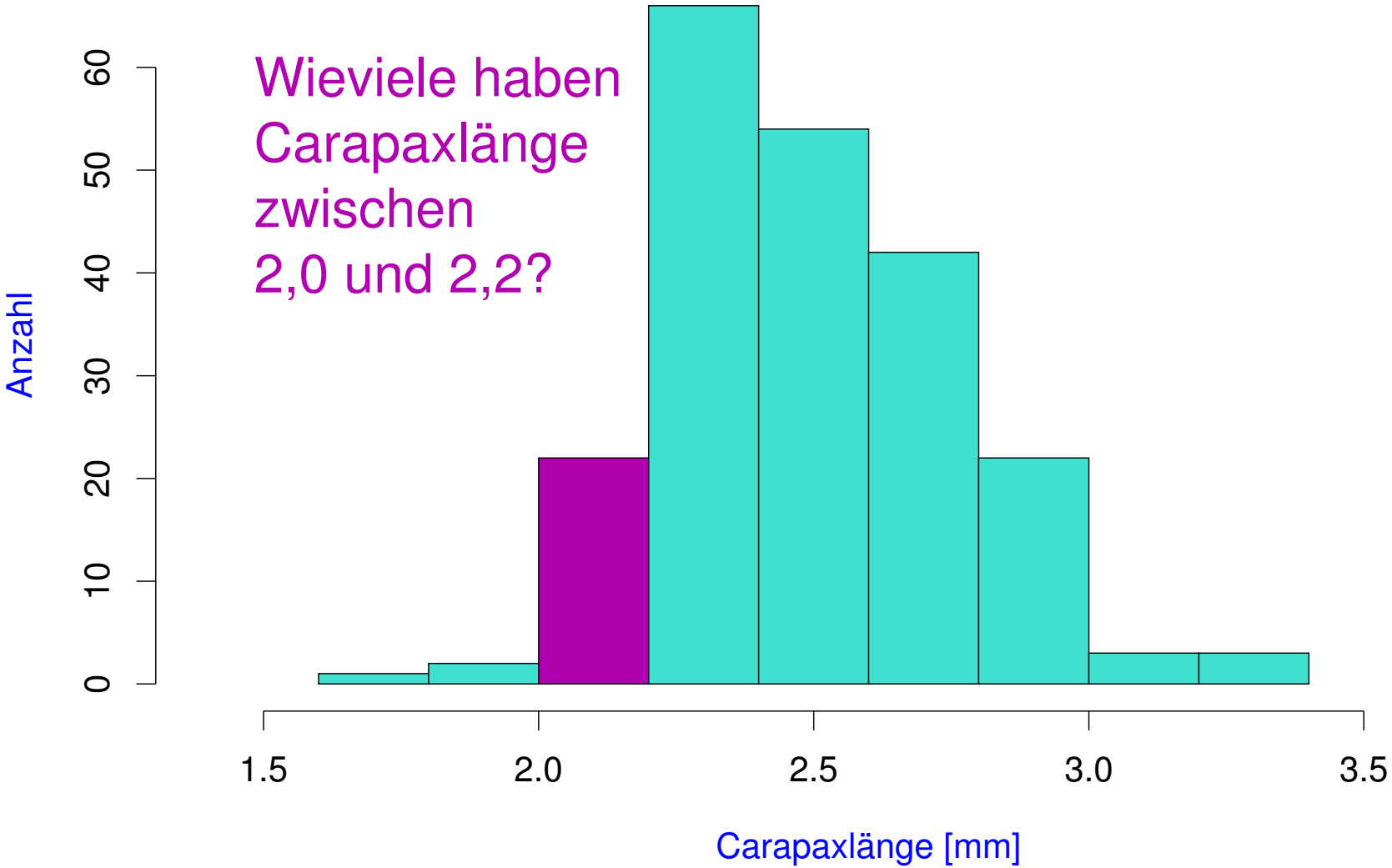
das Histogramm


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]

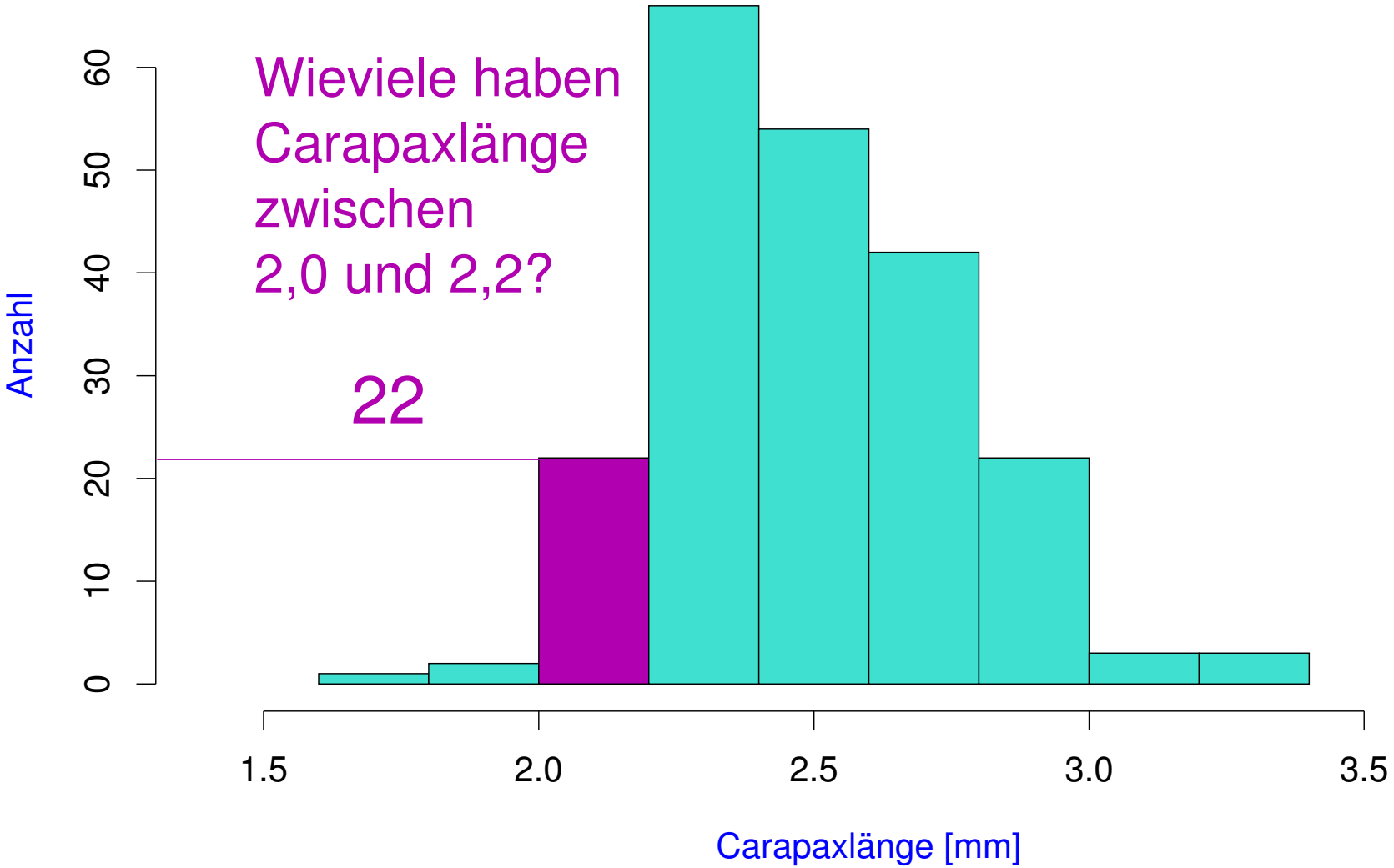
Nichteiertragende Weibchen am 6. Sept. ’88, n=215



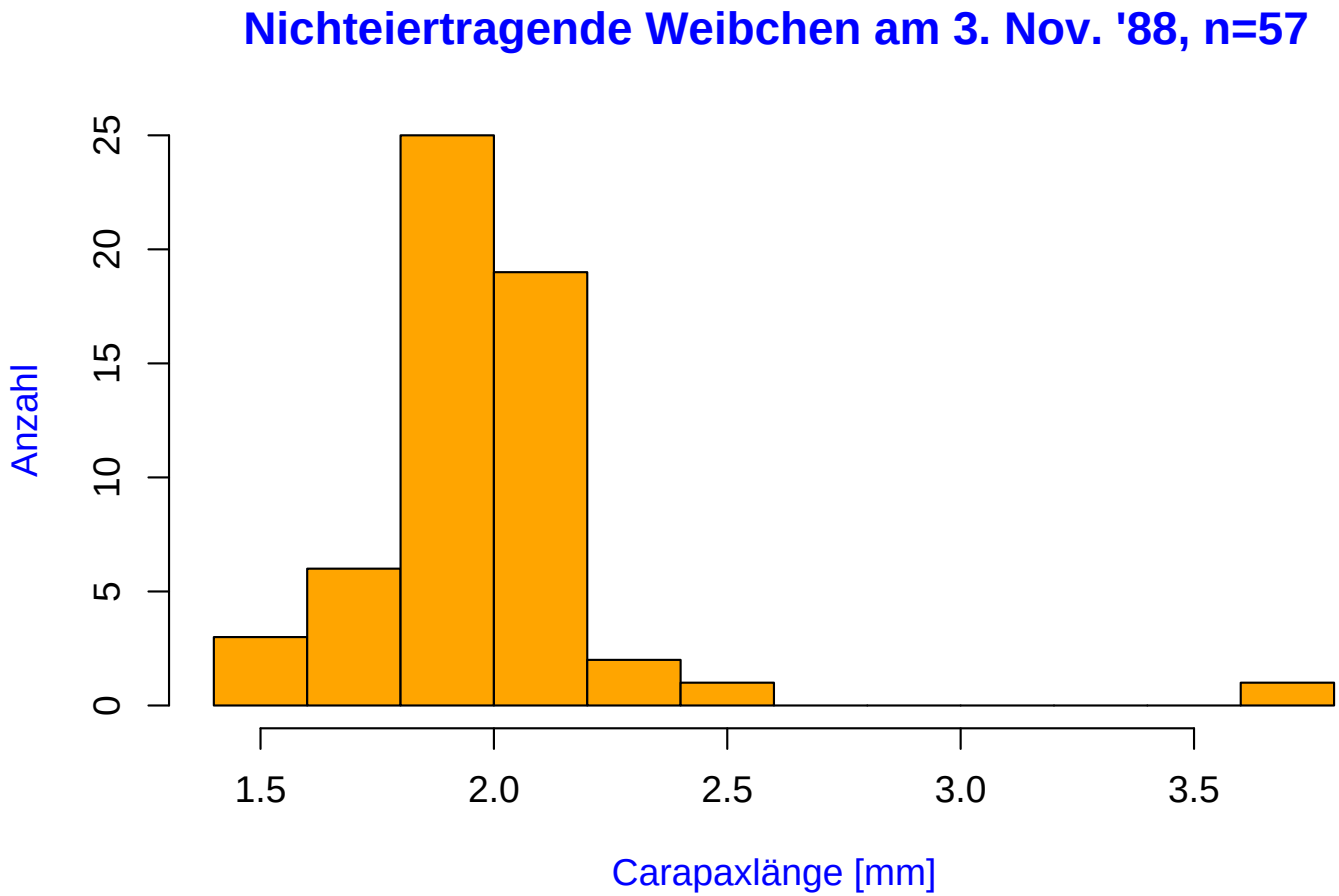
Nichteiertragende Weibchen am 6. Sept. ’88, n=215



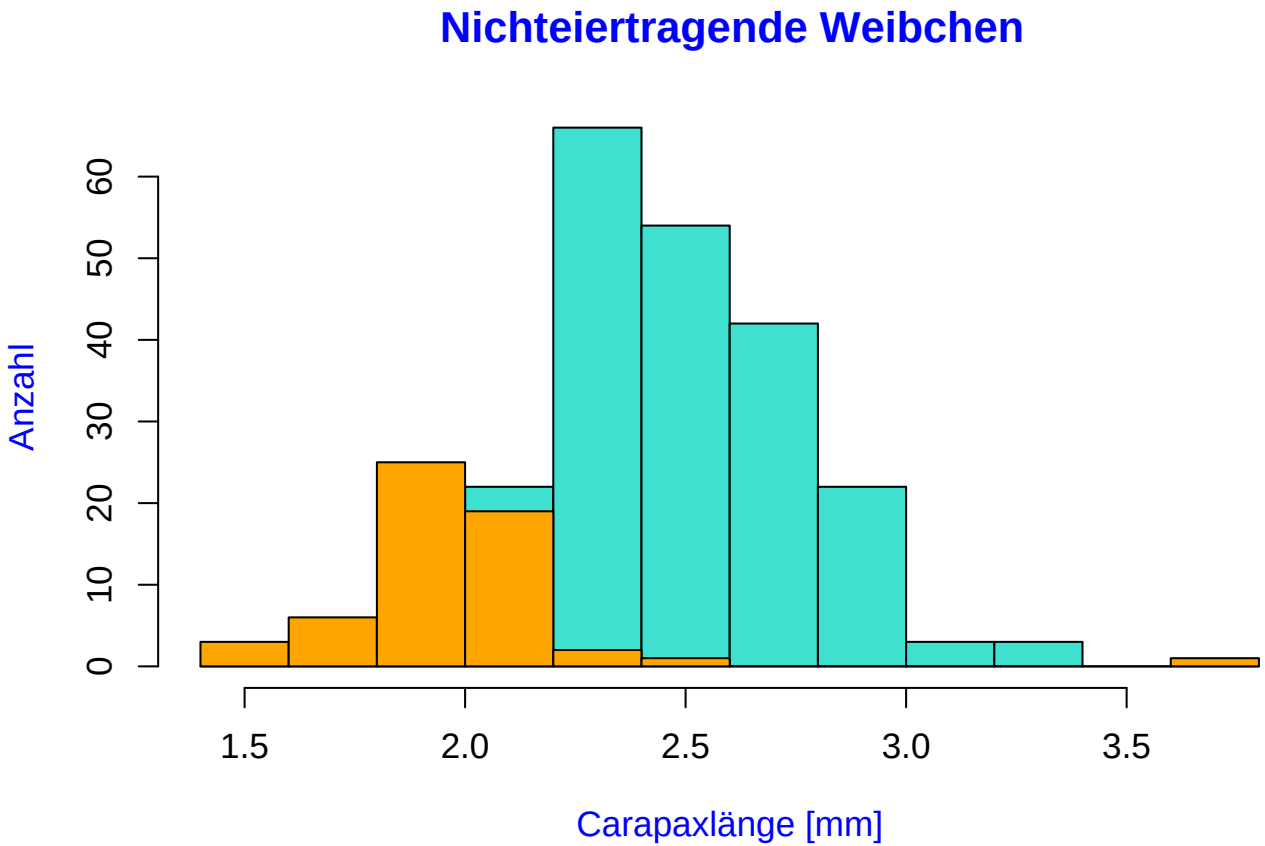
Nichteiertragende Weibchen am 6. Sept. ’88, n=215



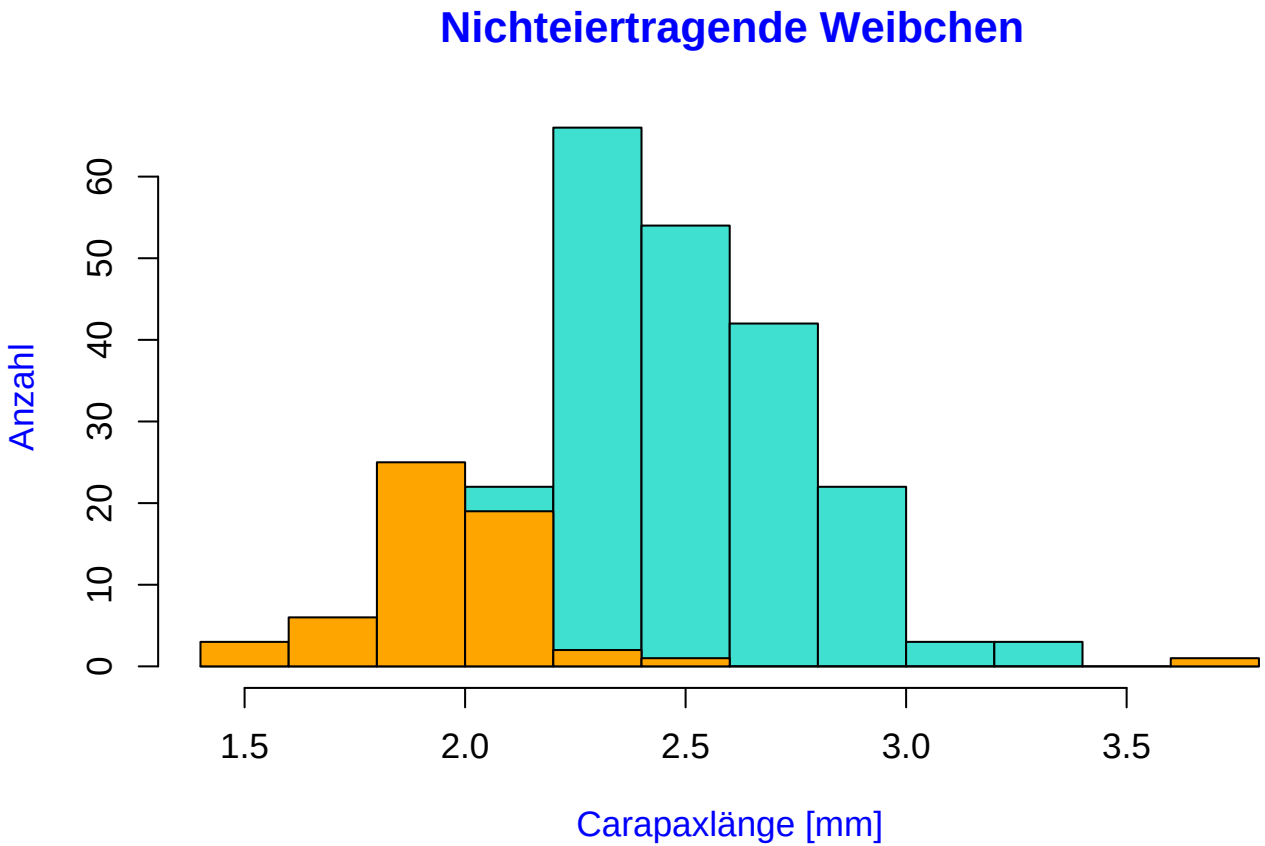
Analoge Daten zwei Monate später (3.11.88):


Nichteiertragende Weibchen am 3. Nov. '88, n=57
Anzahl
0
5
10
15
20
25
1.5
2.0
2.5
3.0
3.5
Carapaxlänge [mm]

Vergleich der beiden Verteilungen



Vergleich der beiden Verteilungen

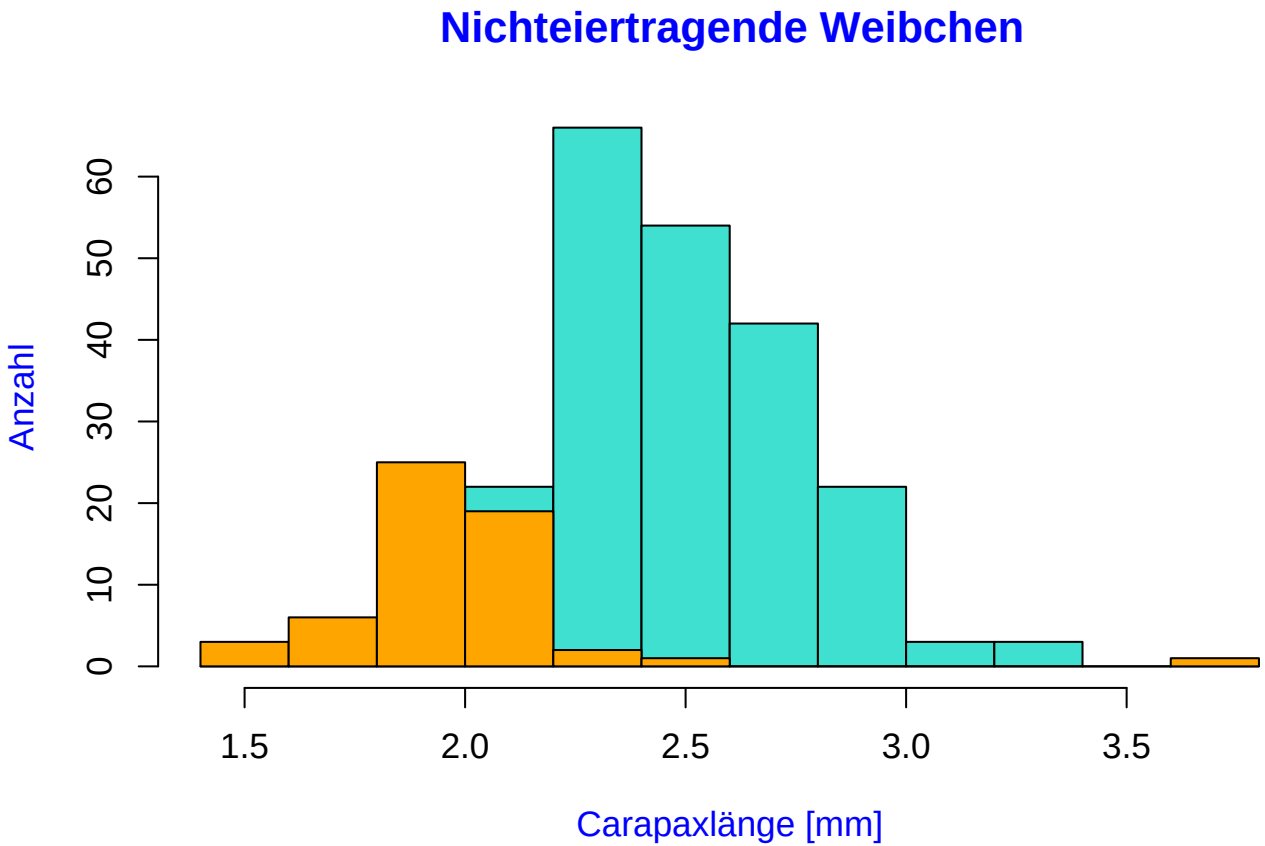


Problem: ungleiche Stichprobenumfänge:

6.Sept: $n = 215$

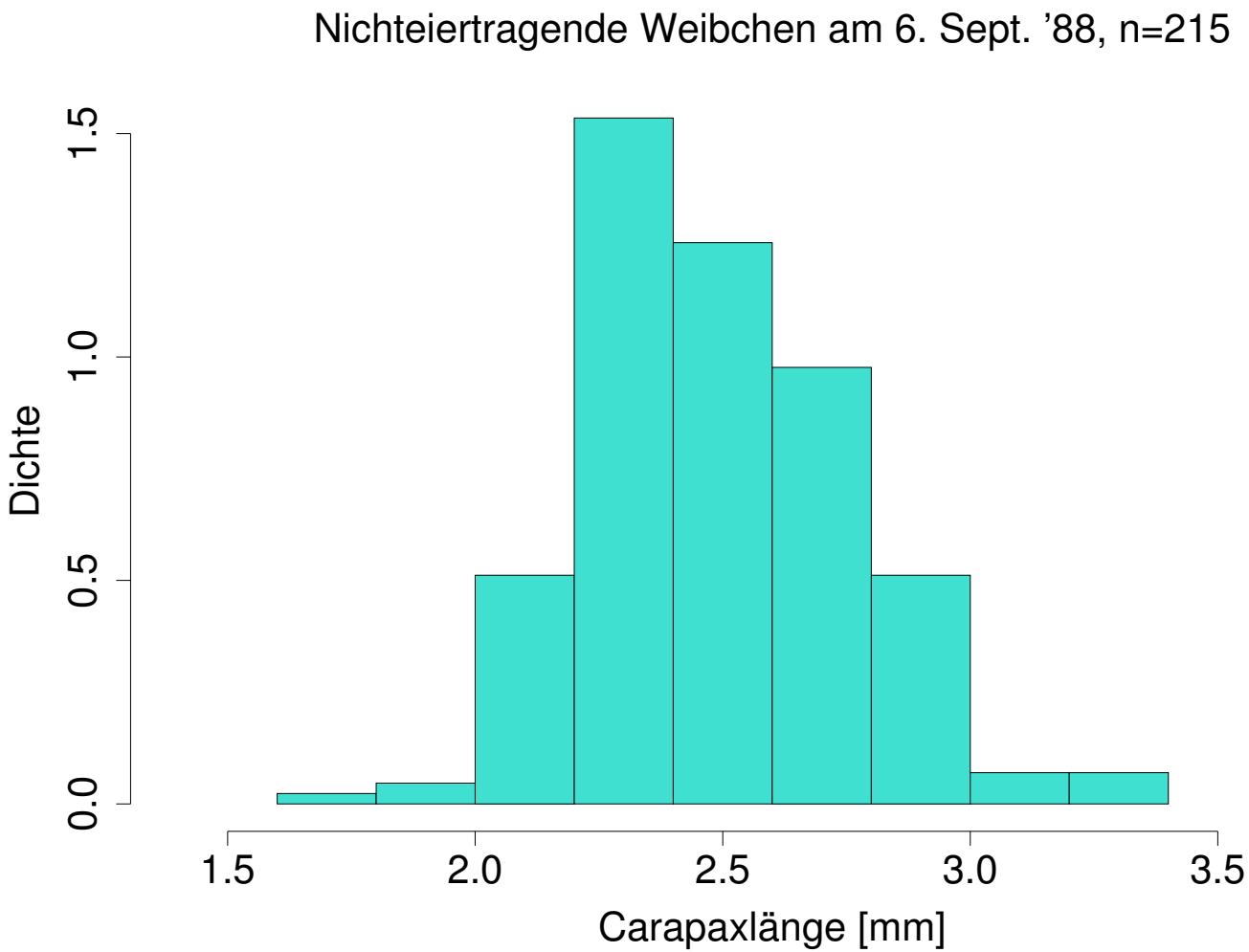
3.Nov : $n = 57$

Vergleich der beiden Verteilungen

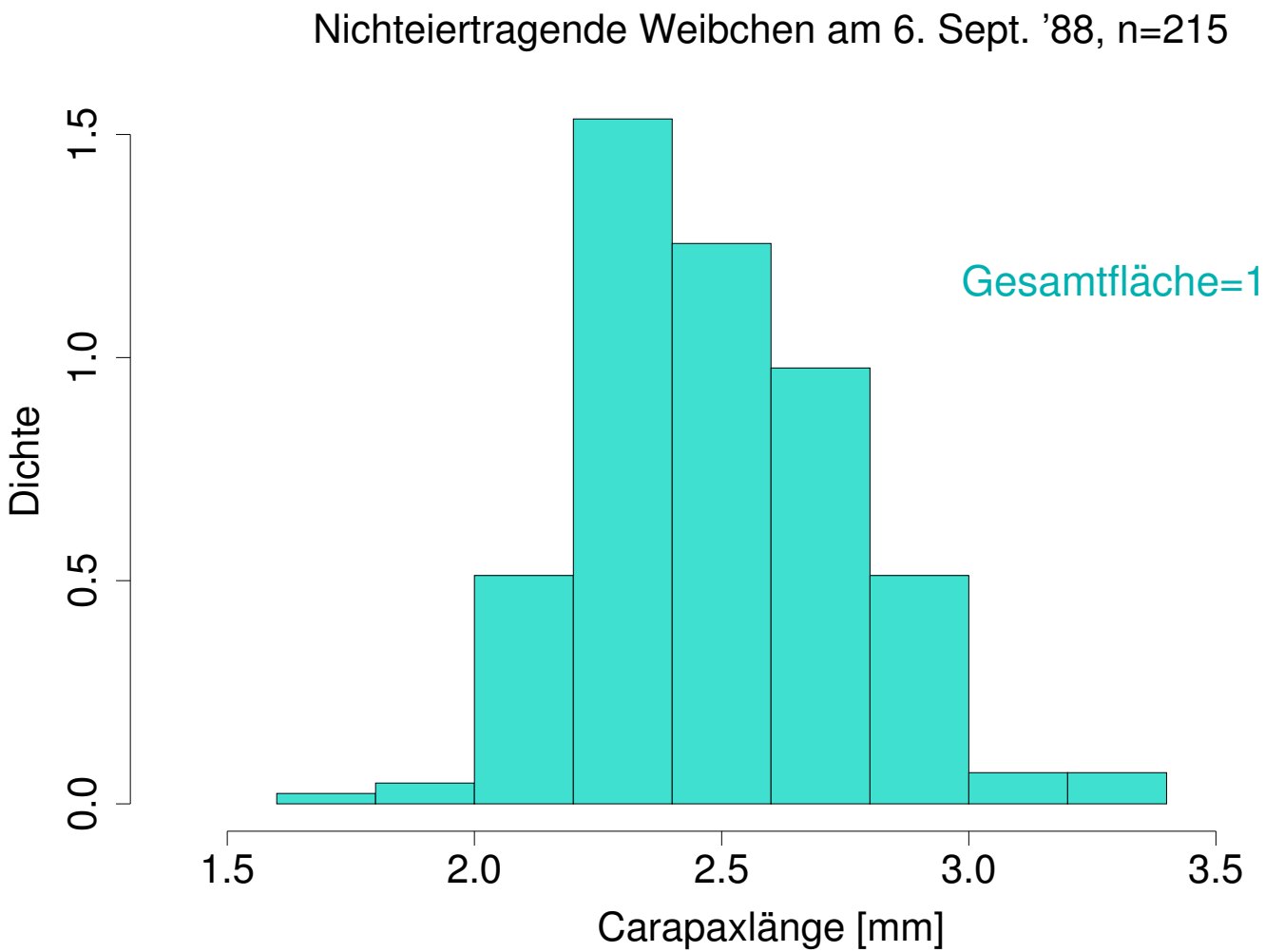


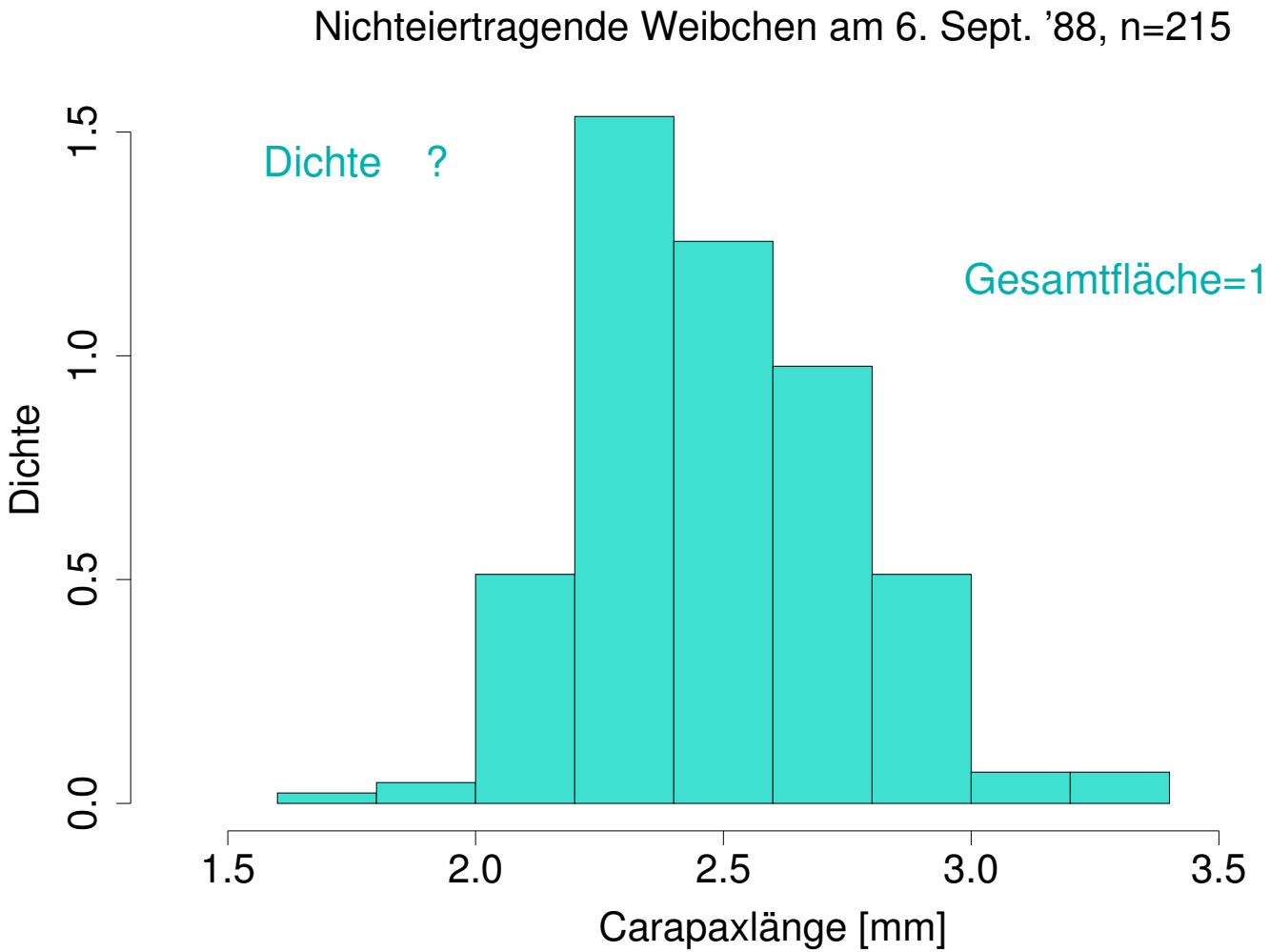
Problem: ungleiche Stichprobenumfänge: 6.Sept: $n = 215$, 3.Nov : $n = 57$

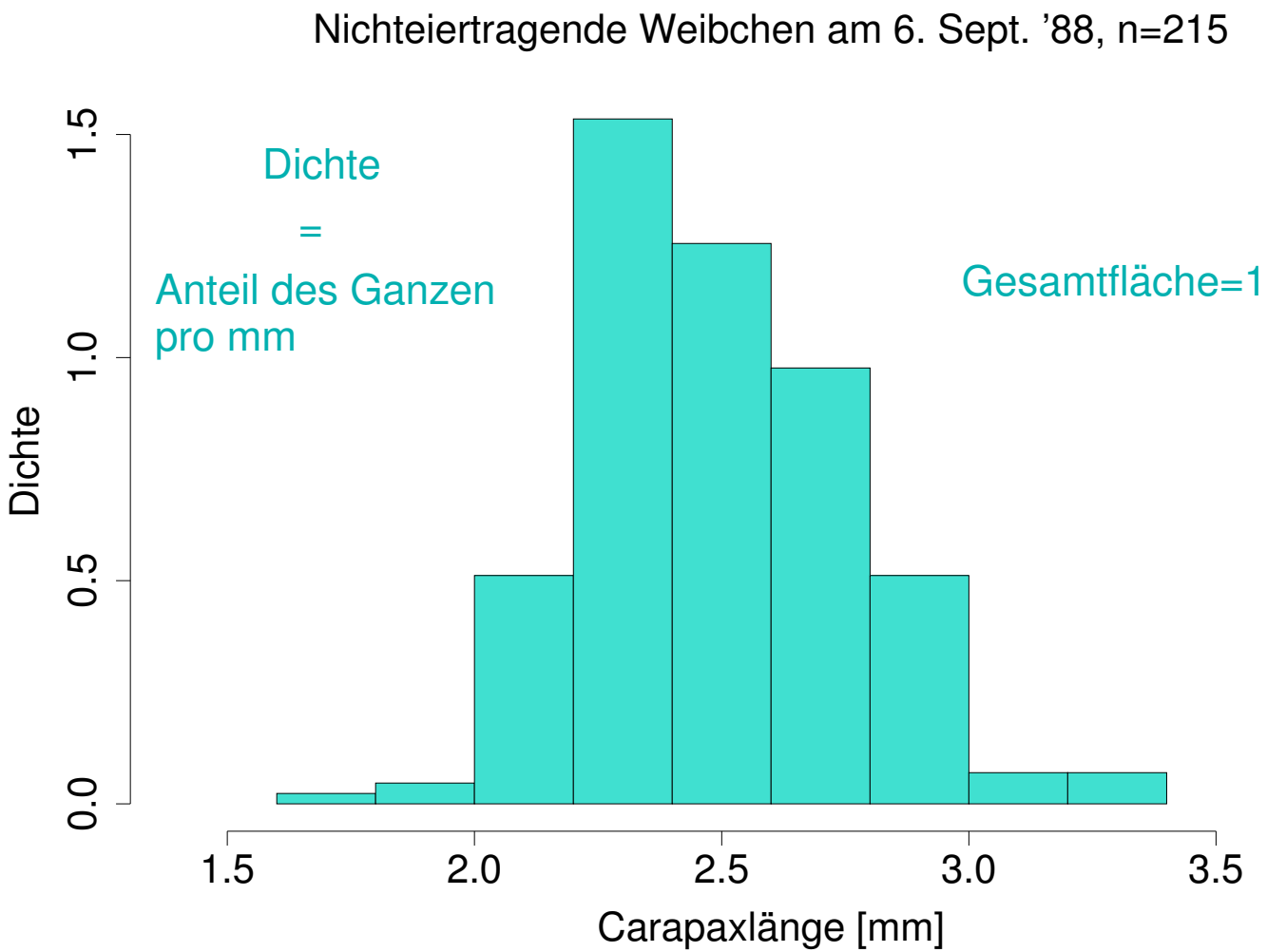
Idee: stauche vertikale Achse so, dass Gesamtfläche = 1.

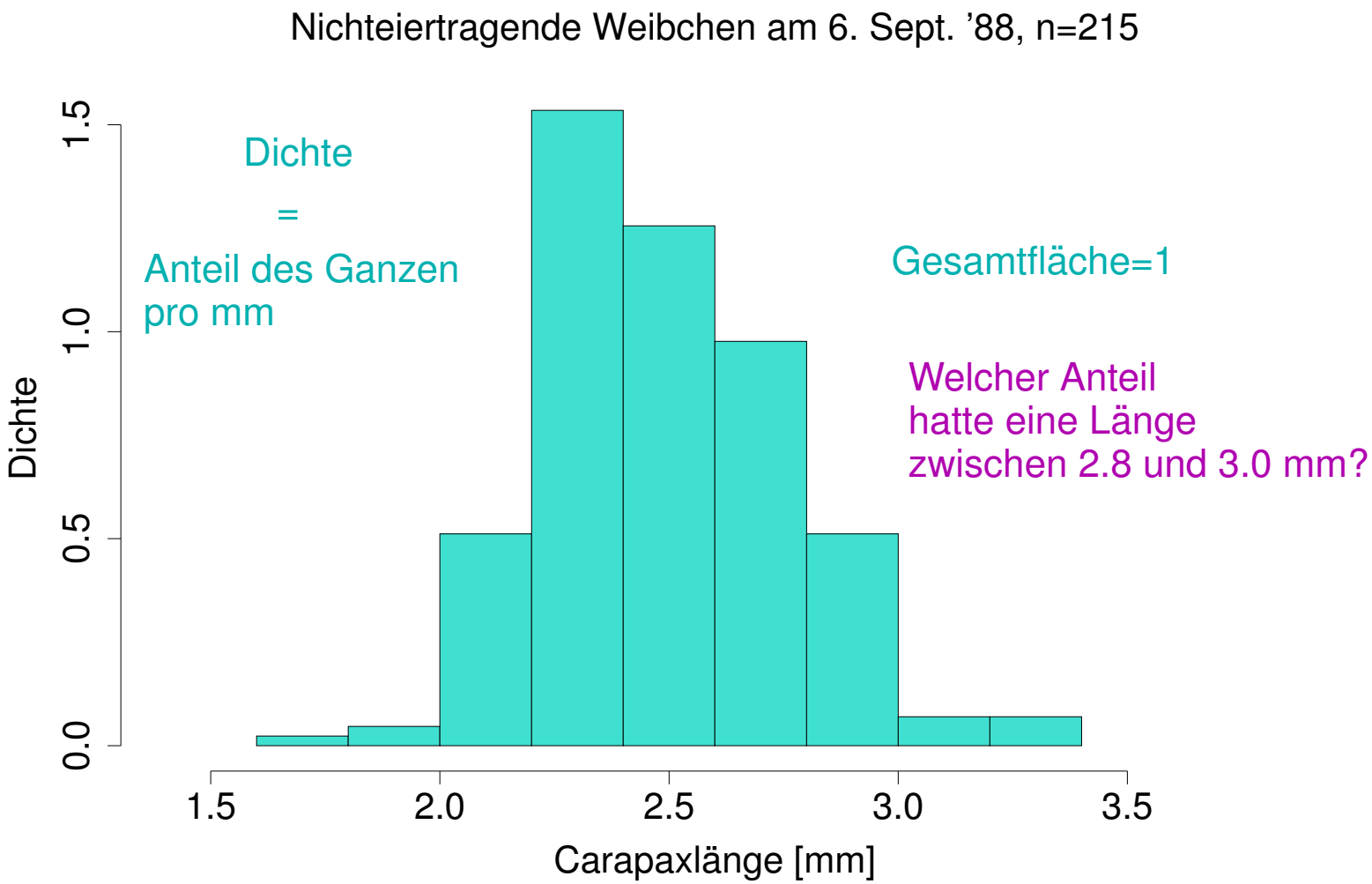

Nichteiertragende Weibchen am 6. Sept. ’88, n=215
Dichte
0.0
0.5
1.0
1.5
1.5
2.0
2.5
3.0
3.5
Carapaxlänge [mm]

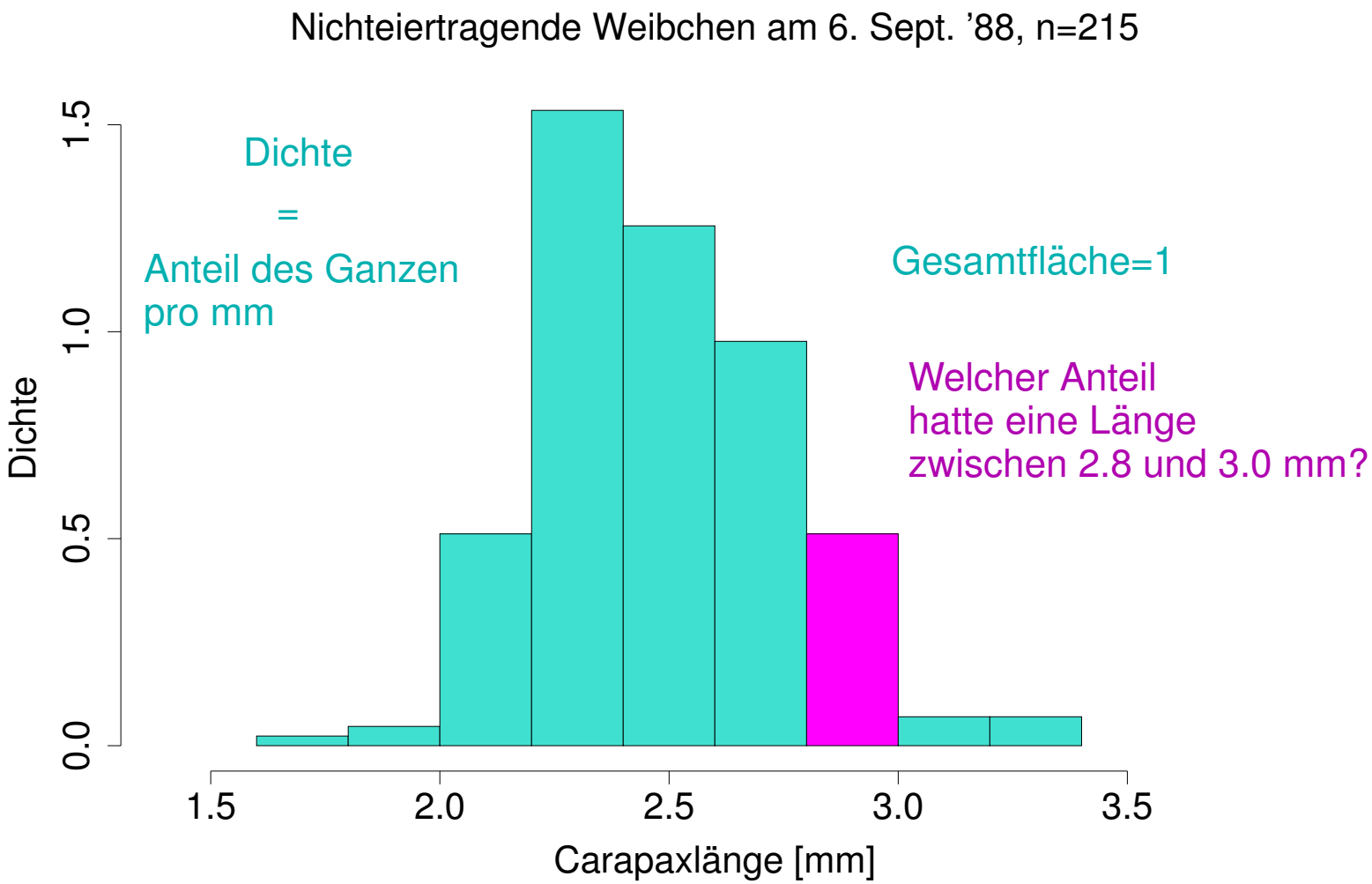
Die neue
vertikale Koordinate
ist jetzt eine
Dichte
(engl. density).

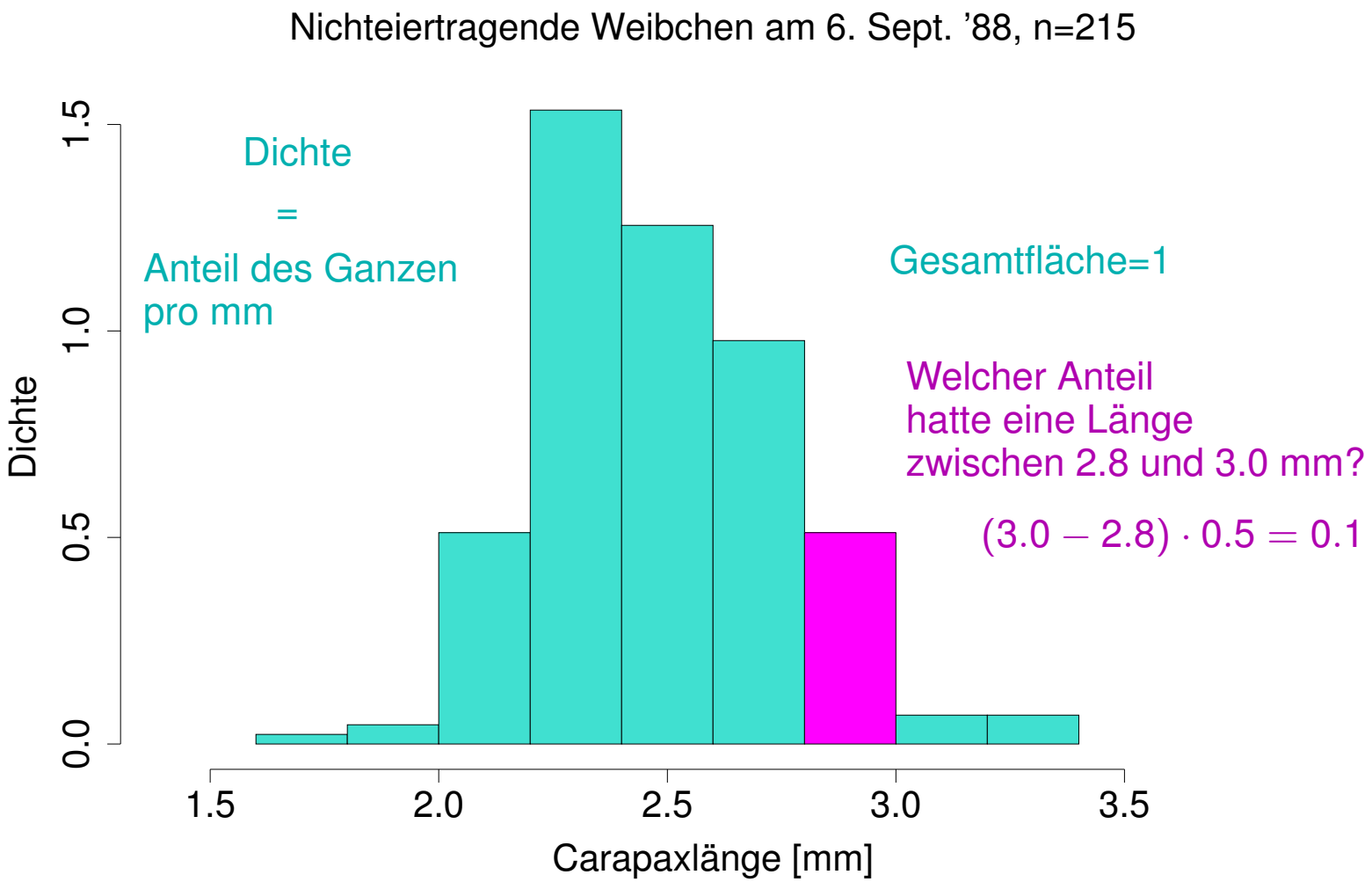

Nichteiertragende Weibchen am 6. Sept. ’88, n=215
Gesamtfläche=1
Dichte
0.0
0.5
1.0
1.5
1.5
2.0
2.5
3.0
3.5
Carapaxlänge [mm]

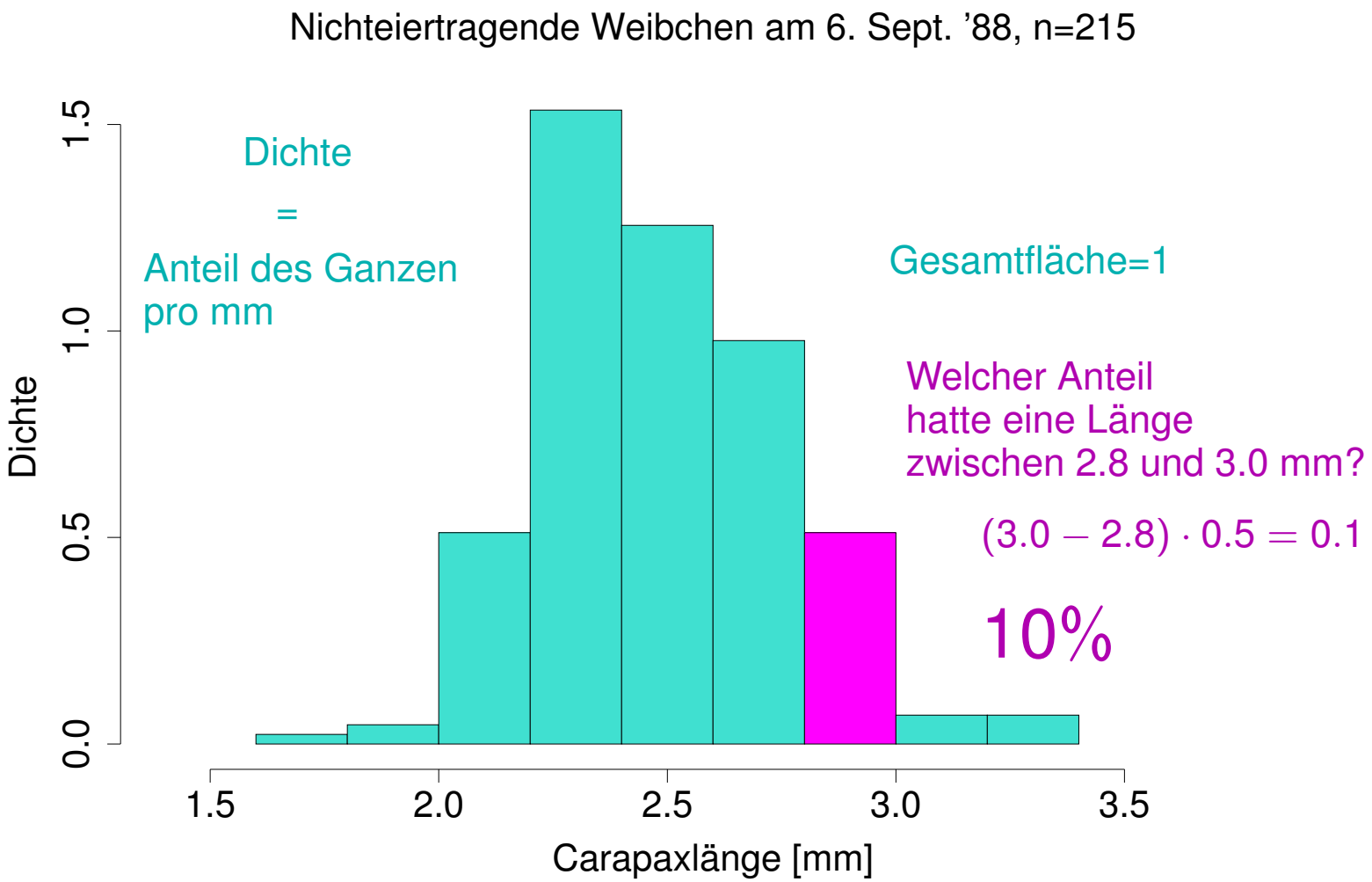

Nichteiertragende Weibchen am 6. Sept. ’88, n=215
Dichte ?
Gesamtfläche=1
Dichte
0.0
0.5
1.0
1.5
1.5
2.0
2.5
3.0
3.5
Carapaxlänge [mm]


Nichteiertragende Weibchen am 6. Sept. ’88, n=215
Dichte
=
Anteil des Ganzen
pro mm
Gesamtfläche=1
Dichte
0.0
0.5
1.0
1.5
1.5
2.0
2.5
3.0
3.5
Carapaxlänge [mm]


Nichteiertragende Weibchen am 6. Sept. '88, n=215
Dichte
=
Anteil des Ganzen
pro mm
Gesamtfläche=1
Welcher Anteil
hatte eine Länge
zwischen 2.8 und 3.0 mm?
Dichte
0.0
0.5
1.0
1.5
1.5
2.0
2.5
3.0
3.5
Carapaxlänge [mm]


Nichteiertragende Weibchen am 6. Sept. '88, n=215
Dichte
=
Anteil des Ganzen
pro mm
Gesamtfläche=1
Welcher Anteil
hatte eine Länge
zwischen 2.8 und 3.0 mm?
Dichte
0.0
0.5
1.0
1.5
1.5
2.0
2.5
3.0
3.5
Carapaxlänge [mm]


Nichteiertragende Weibchen am 6. Sept. '88, n=215
Dichte
=
Anteil des Ganzen
pro mm
Gesamtfläche=1
Welcher Anteil
hatte eine Länge
zwischen 2.8 und 3.0 mm?
$(3.0 - 2.8) \cdot 0.5 = 0.1$
Dichte
0.0
0.5
1.0
1.5
1.5
2.0
2.5
3.0
3.5
Carapaxlänge [mm]

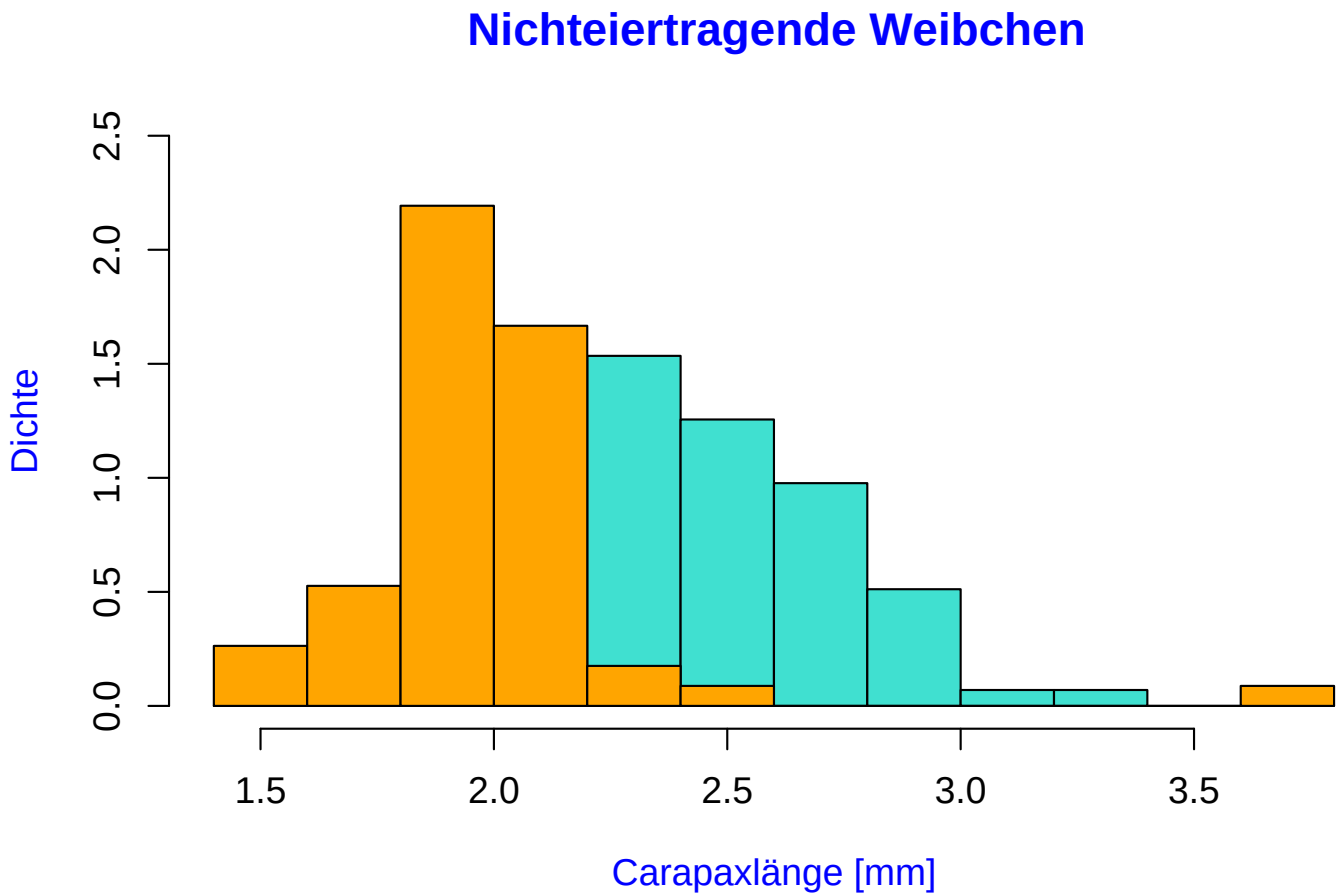

Nichteiertragende Weibchen am 6. Sept. '88, n=215
Dichte
=
Anteil des Ganzen
pro mm
Gesamtfläche=1
Welcher Anteil
hatte eine Länge
zwischen 2.8 und 3.0 mm?
$(3.0 - 2.8) \cdot 0.5 = 0.1$
10%
Dichte
0.0
0.5
1.0
1.5
1.5
2.0
2.5
3.0
3.5
Carapaxlänge [mm]

Die beiden Histogramme sind jetzt vergleichbar

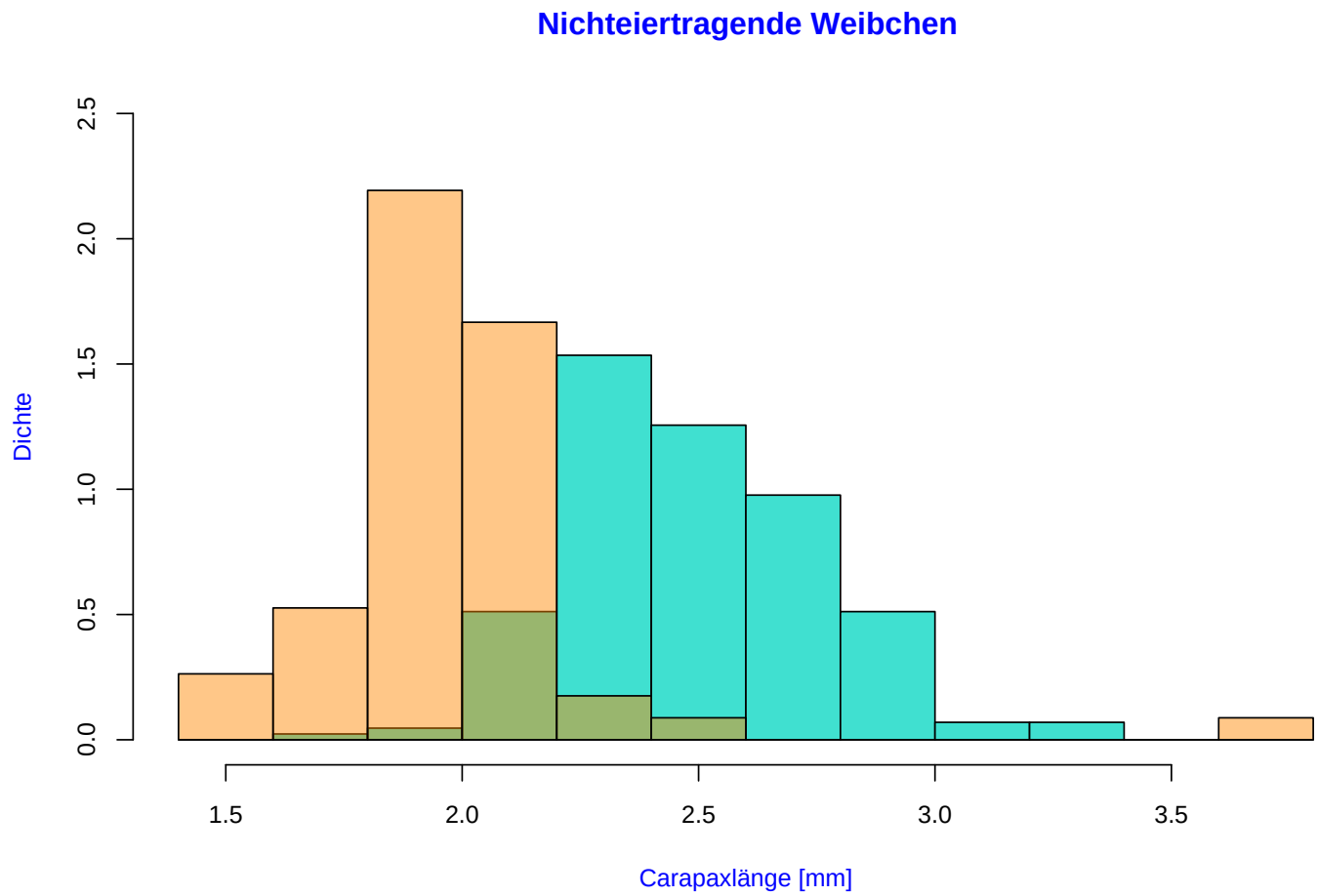
Die beiden Histogramme sind jetzt vergleichbar

(sie haben dieselbe Gesamtfläche).

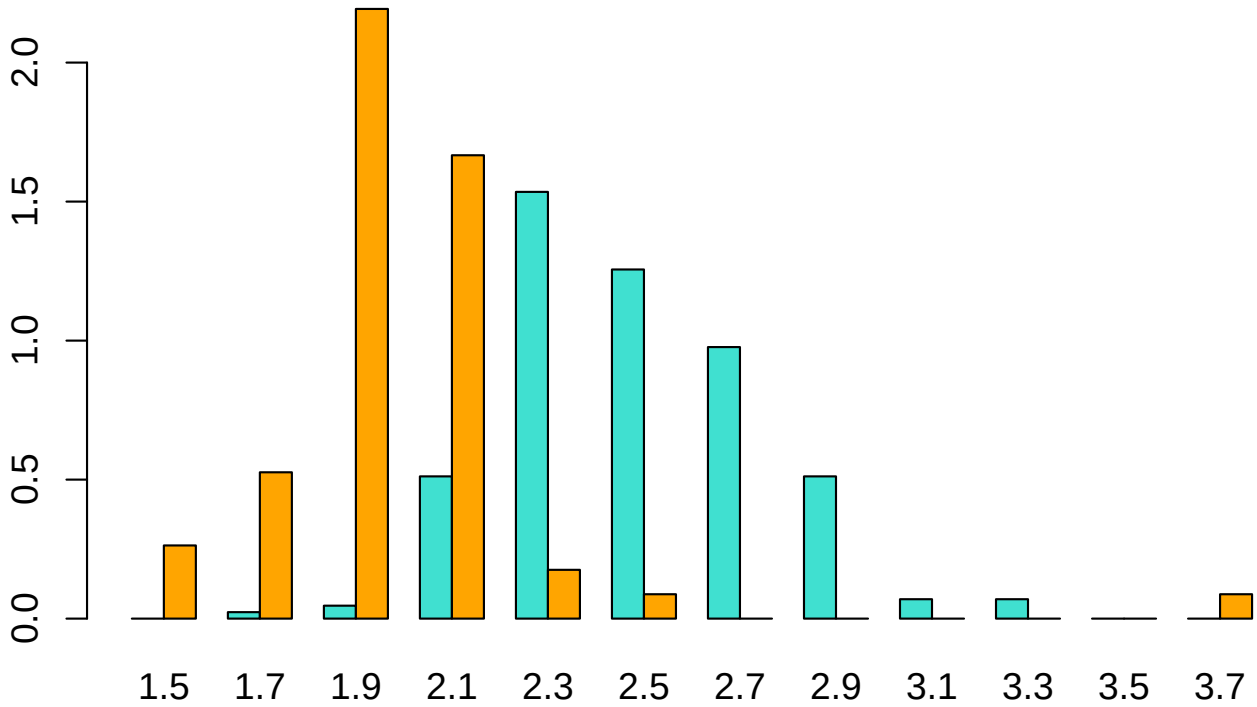
Versuche, die Histogramme zusammen zu zeigen:



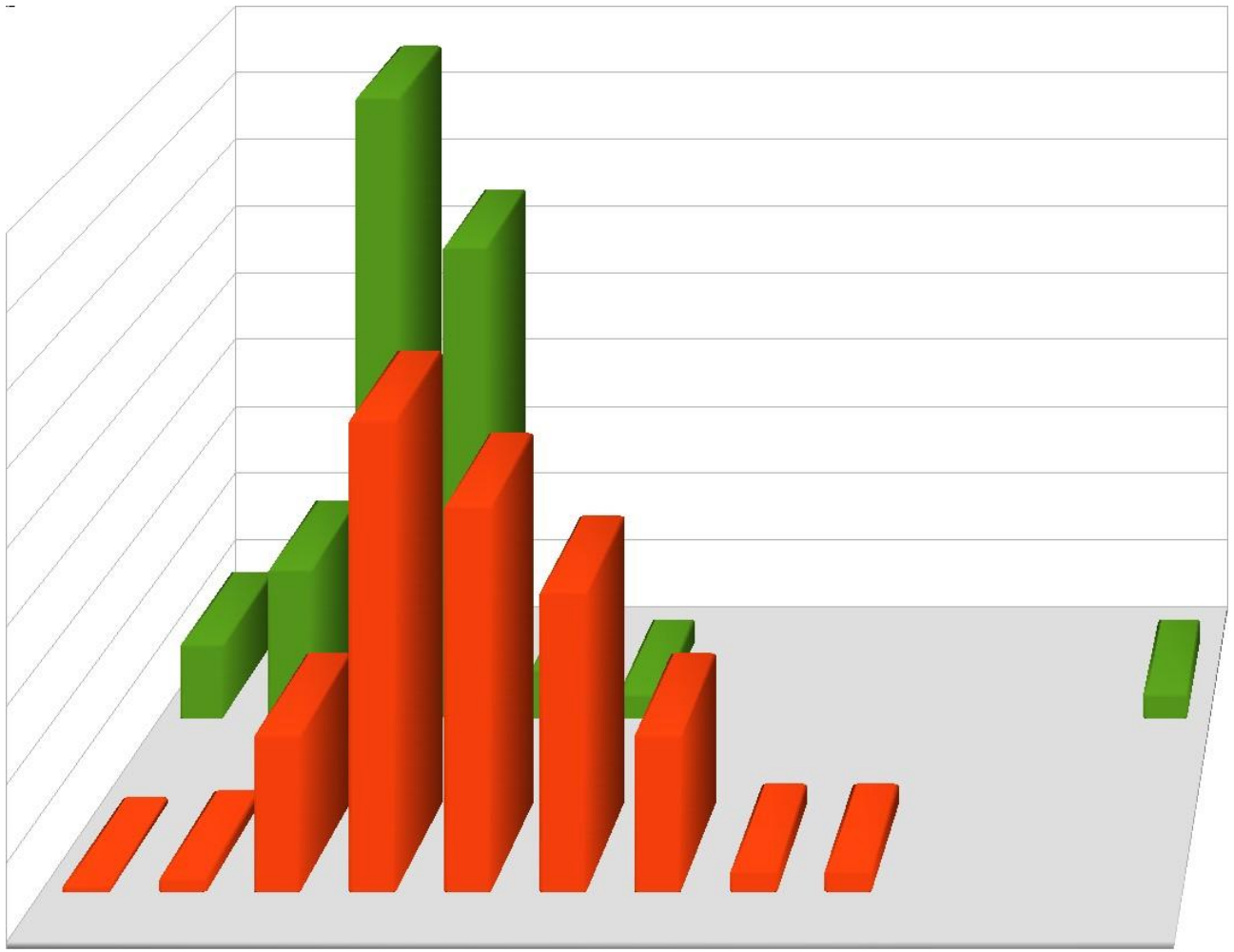
Versuche, die Histogramme zusammen zu zeigen:



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

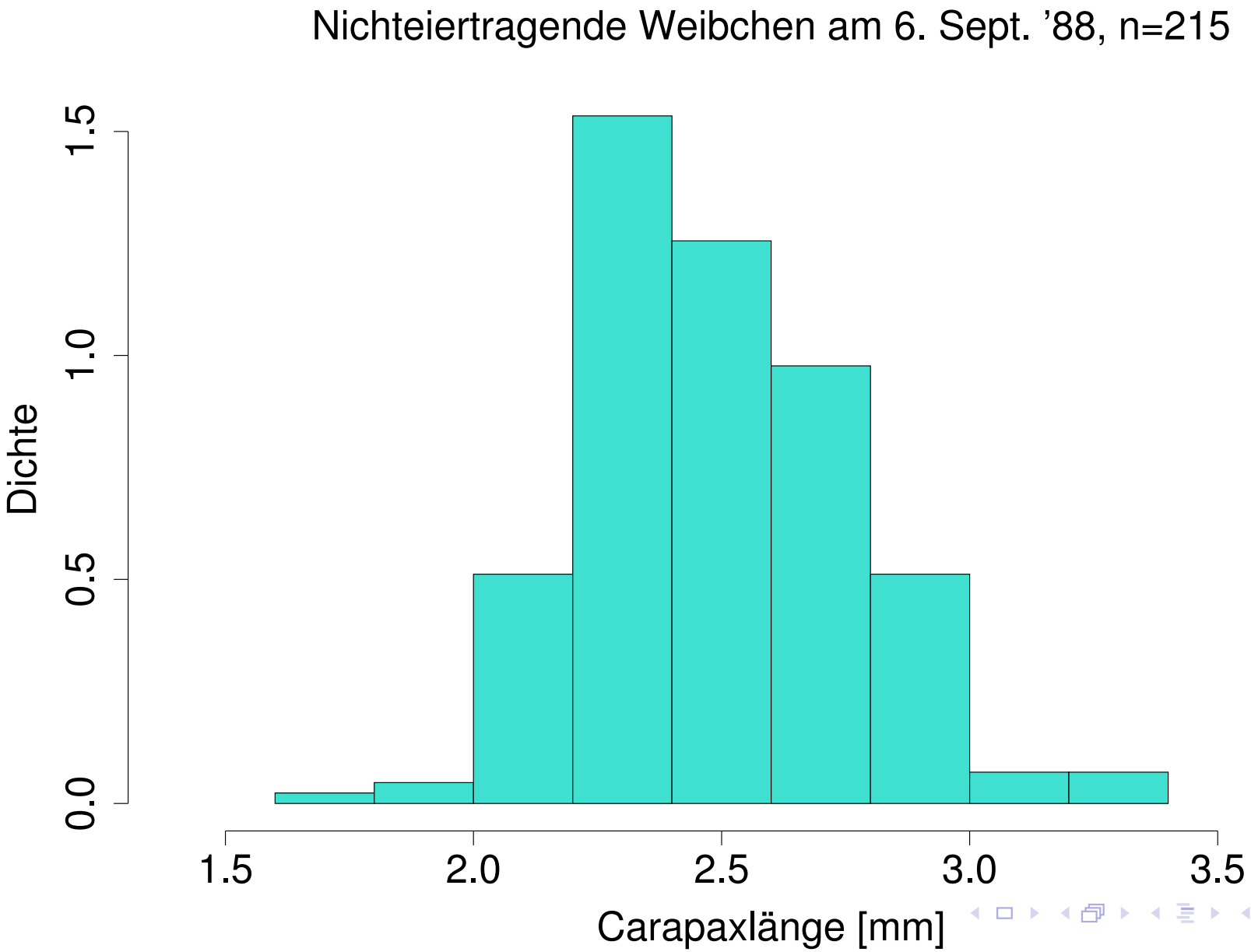
Histogramme kann man nicht ohne weiteres in demselben Graphen darstellen,

Problem

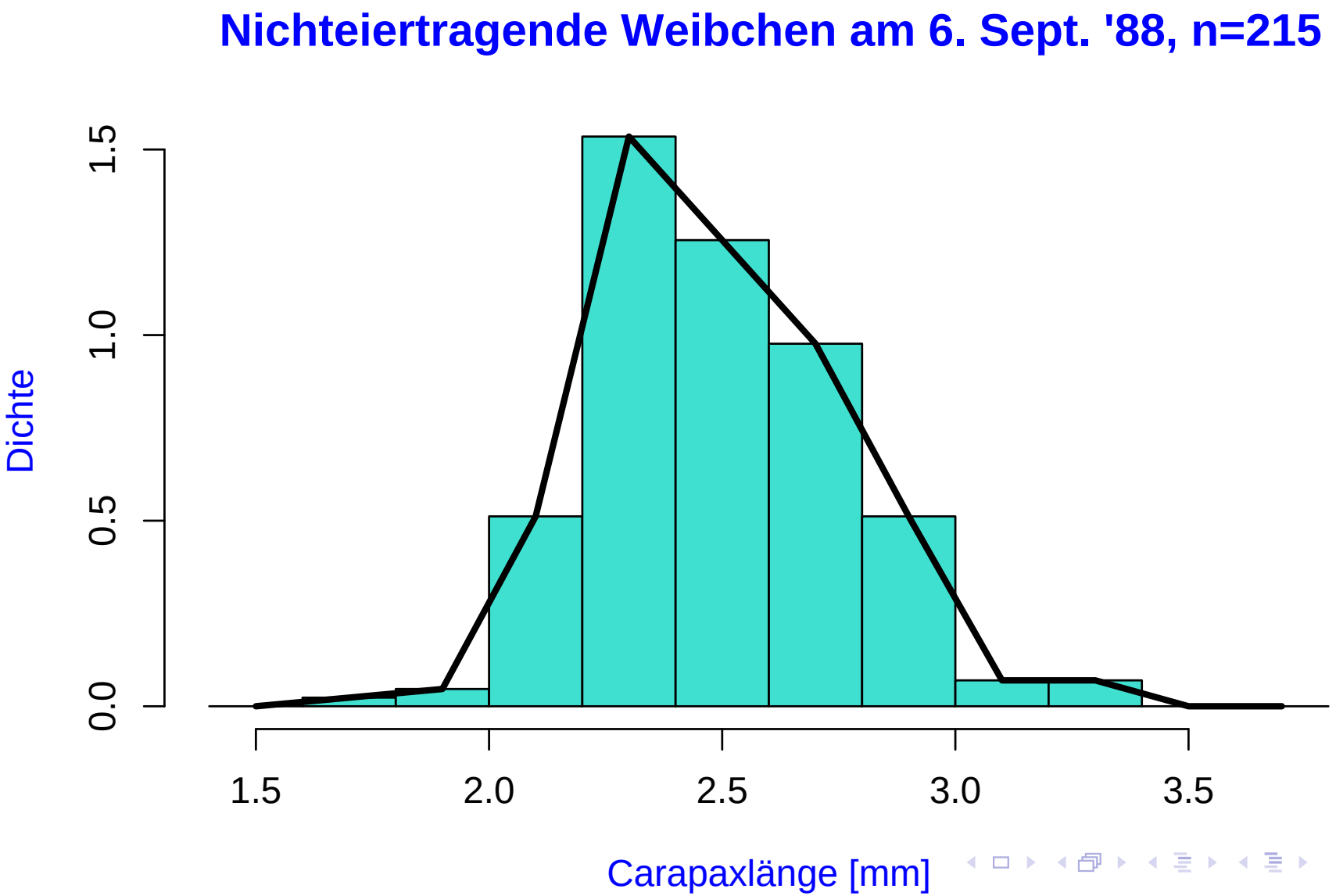
Histogramme kann man nicht ohne weiteres in demselben Graphen darstellen,

weil sie einander überdecken würden.

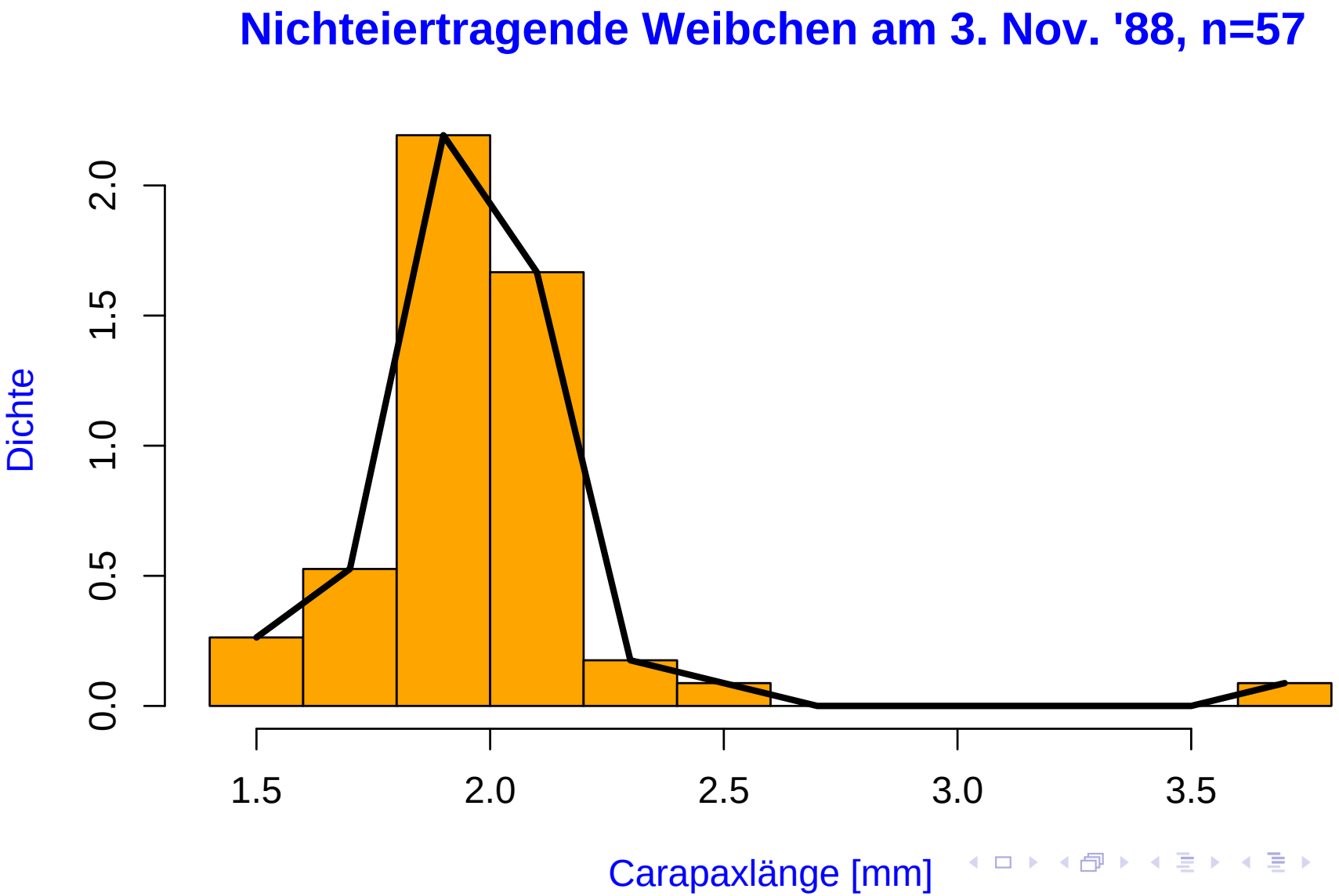
Einfache und klare Lösung: Dichtepolygone



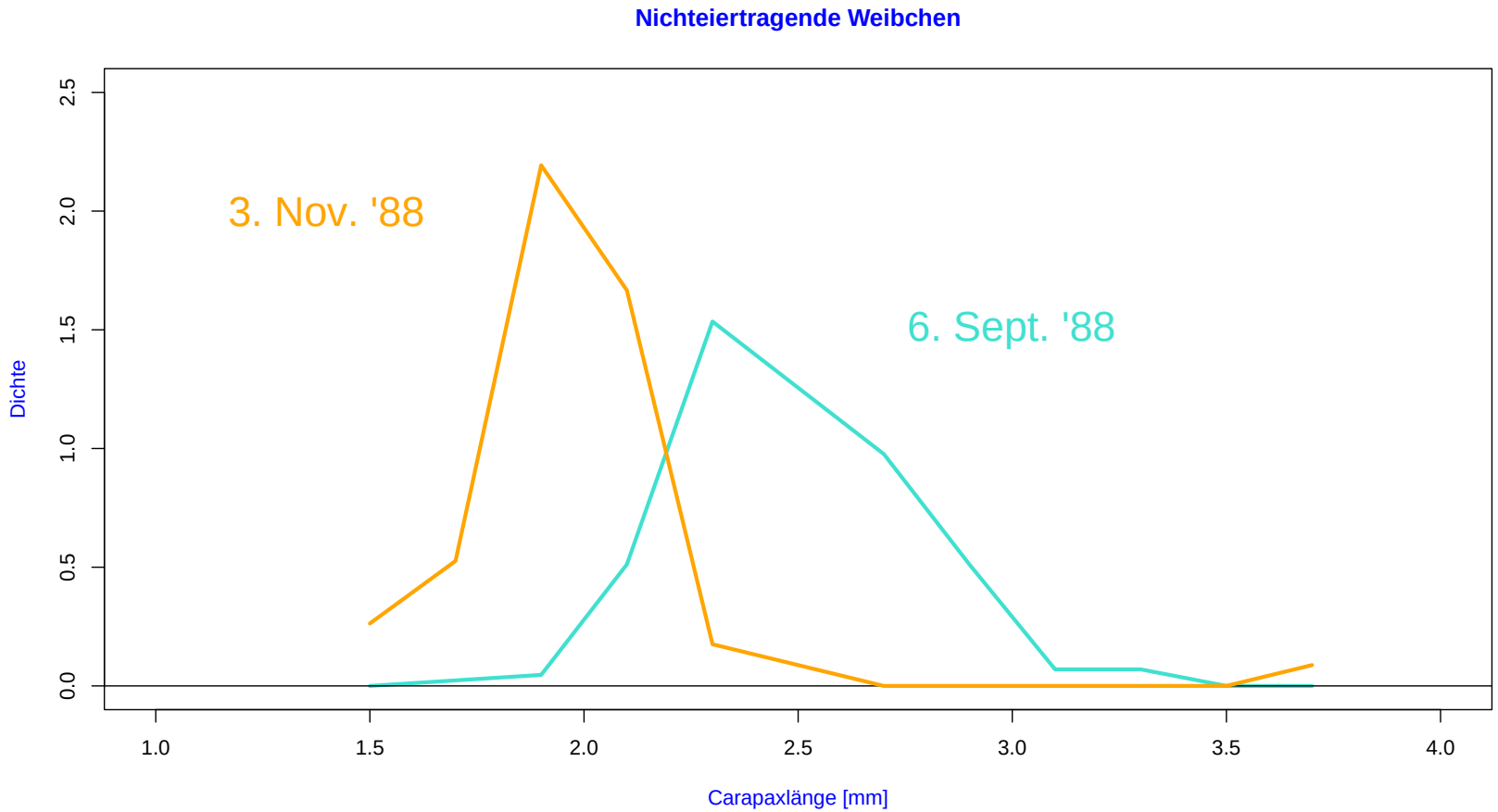
Einfache und klare Lösung: Dichtepolygone



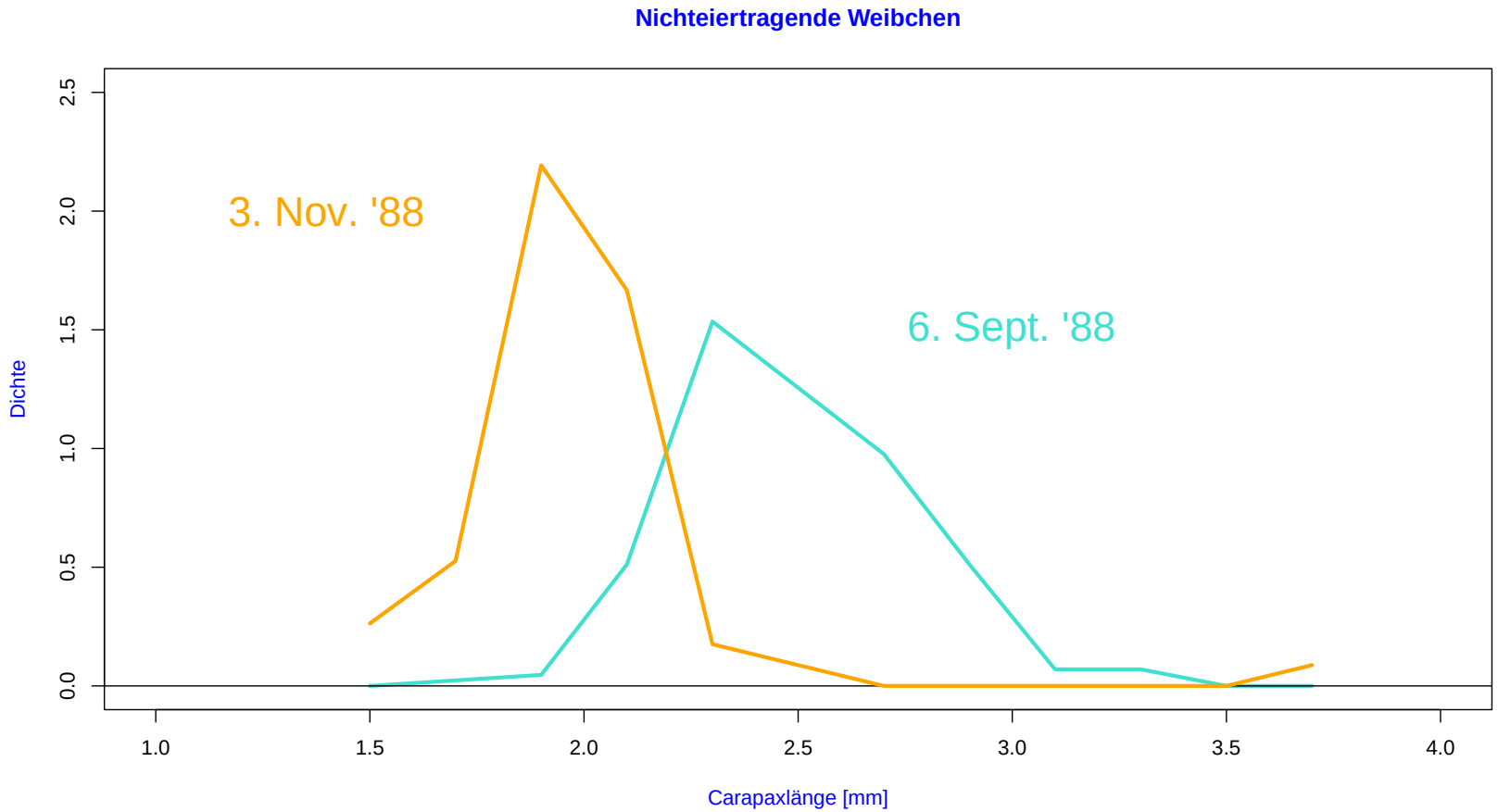
Einfache und klare Lösung: Dichtepolygone



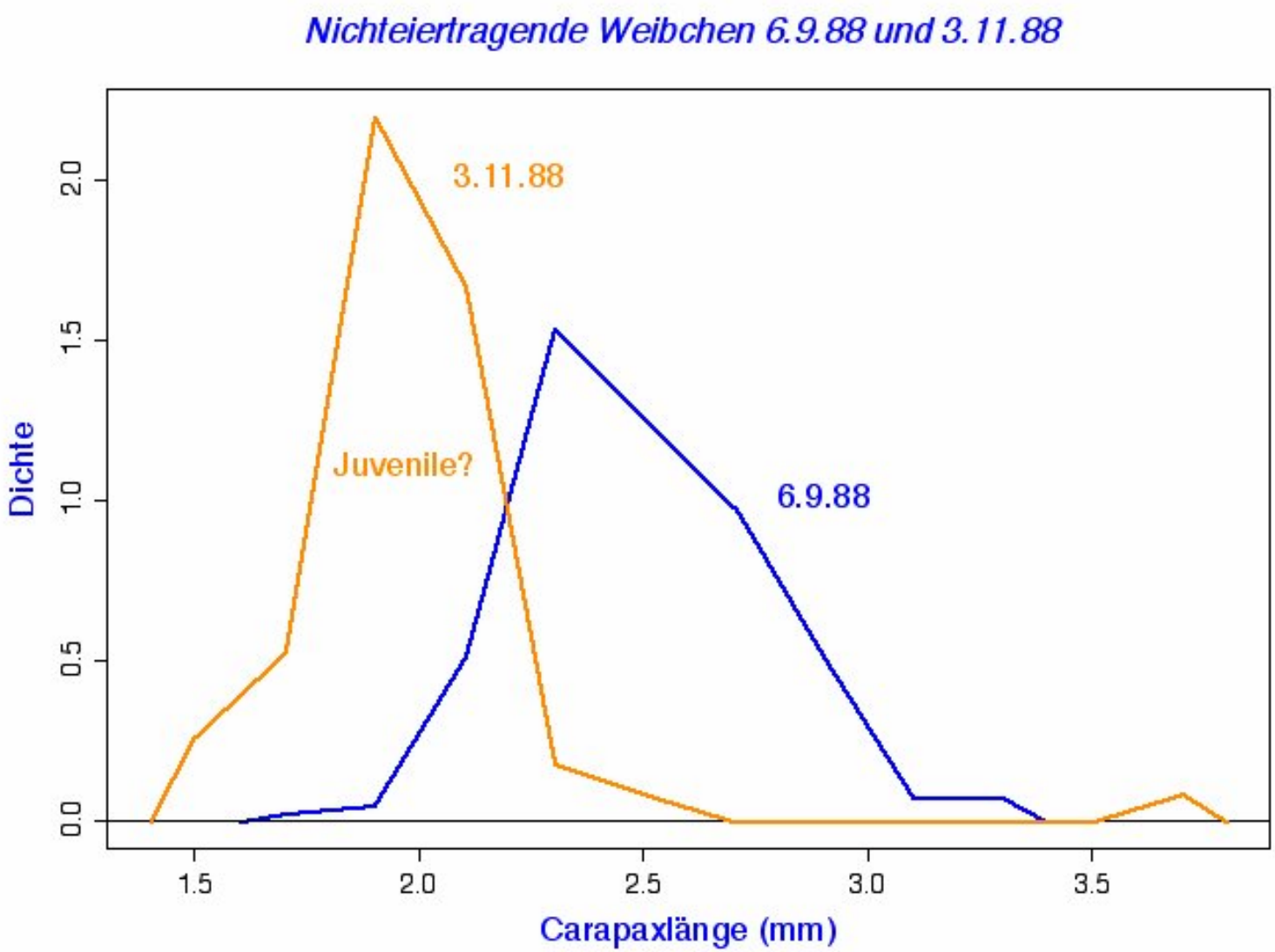
Zwei und mehr Dichtepolygone in einem Plot

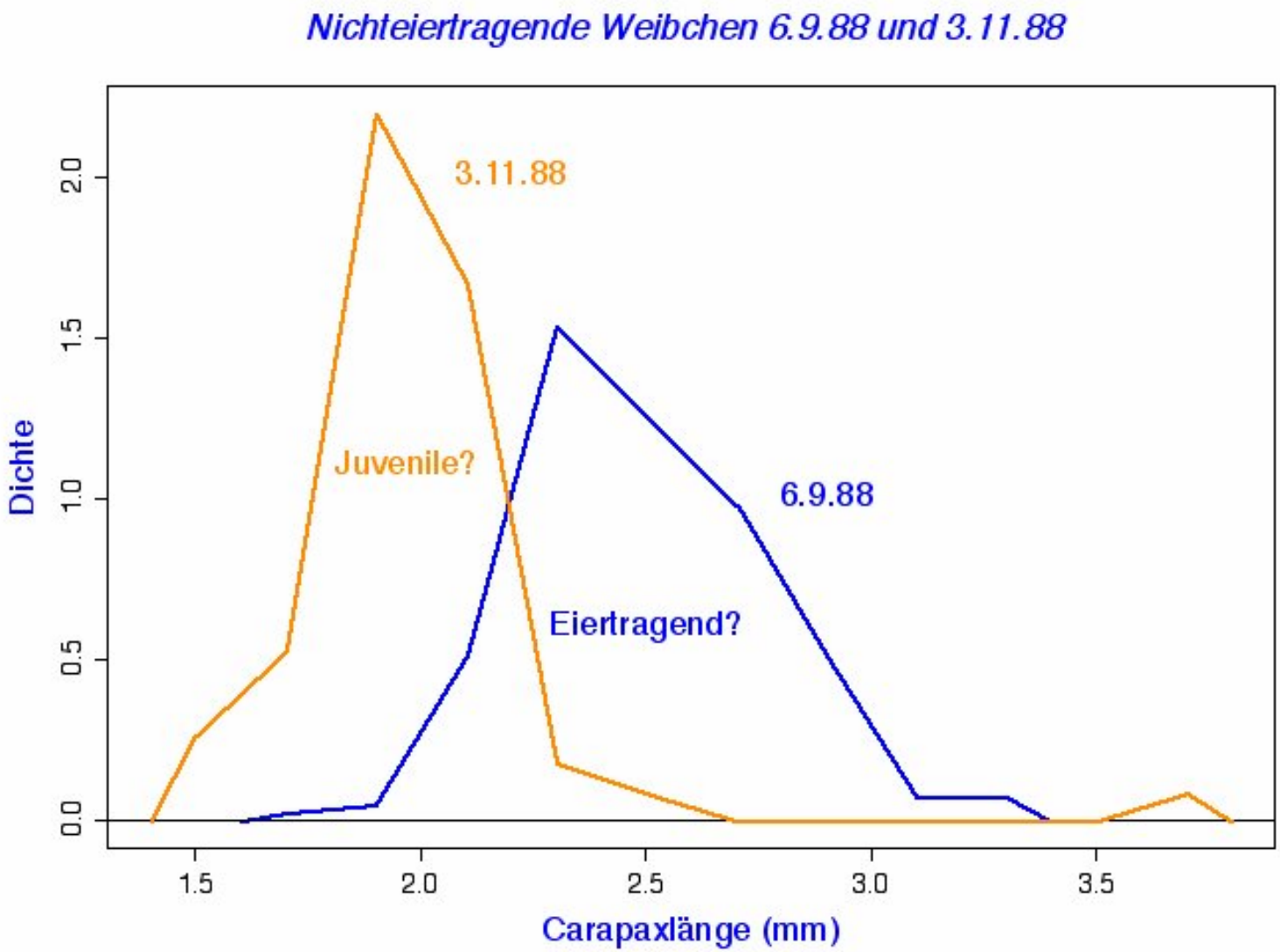


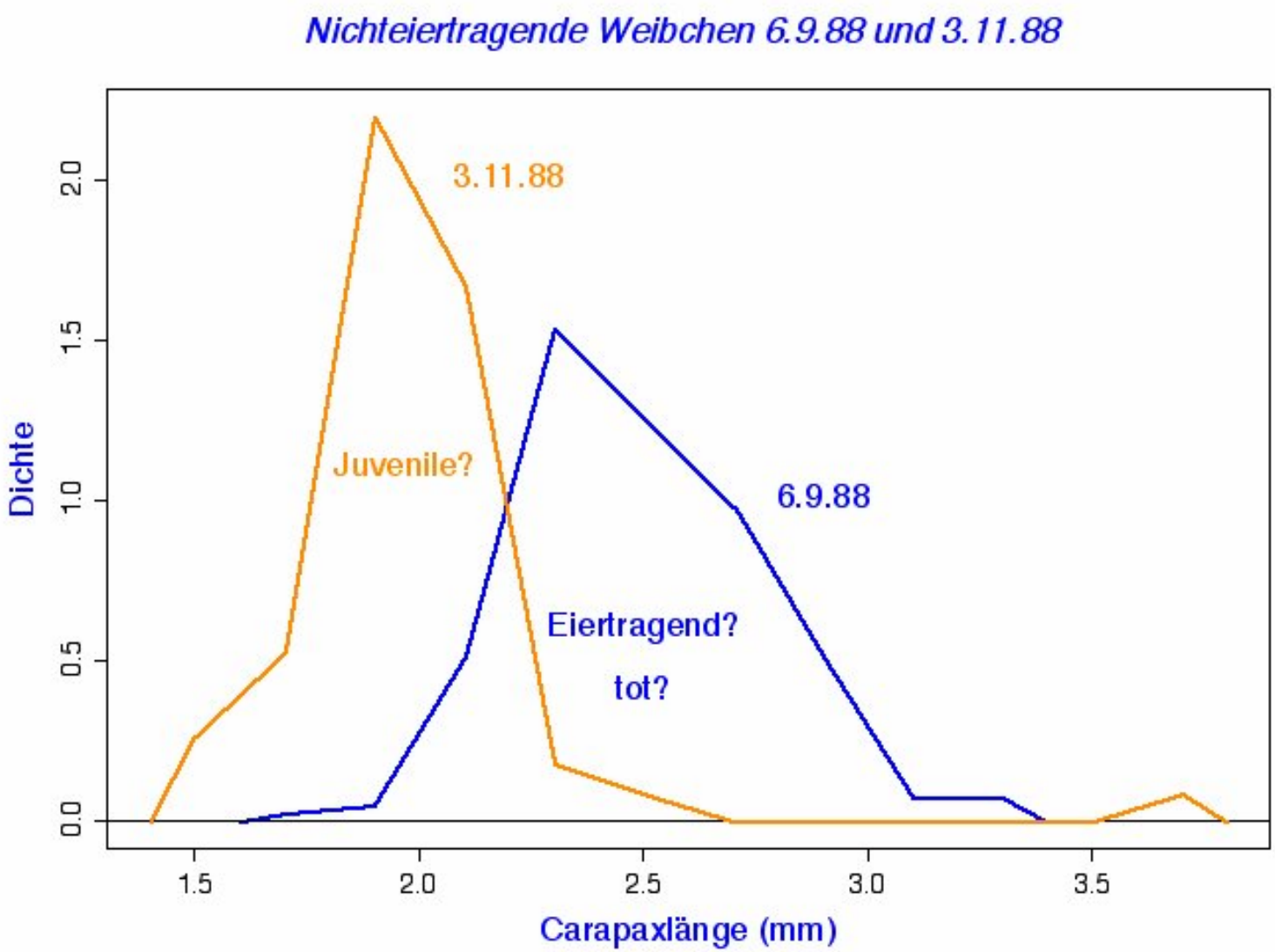
Zwei und mehr Dichtepolygone in einem Plot



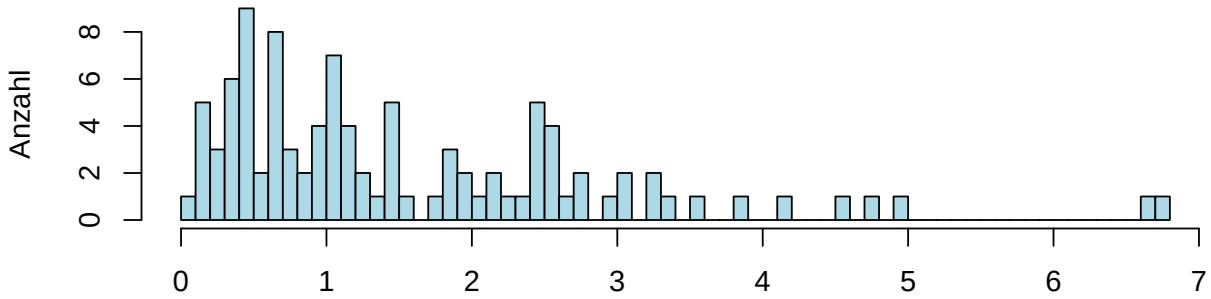
Biologische Interpretation der Verschiebung?


Nichteiertragende Weibchen 6.9.88 und 3.11.88
3.11.88
Juvenile?
6.9.88
Dichte
0.0
0.5
1.0
1.5
2.0
1.5
2.0
2.5
3.0
3.5
Carapaxlänge (mm)

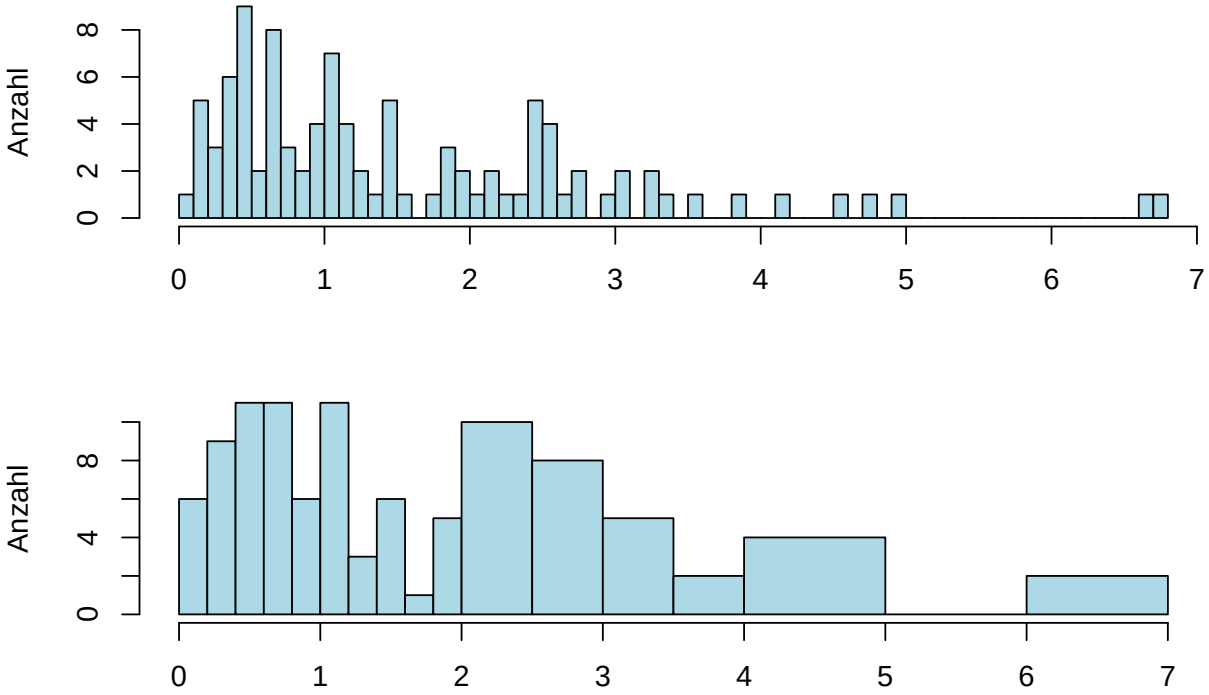

Nichteiertragende Weibchen 6.9.88 und 3.11.88
3.11.88
Juvenile?
6.9.88
Eiertragend?
Dichte
0.0
0.5
1.0
1.5
2.0
1.5
2.0
2.5
3.0
3.5
Carapaxlänge (mm)


Nichteiertragende Weibchen 6.9.88 und 3.11.88
3.11.88
Juvenile?
6.9.88
Eiertragend?
tot?
Dichte
0.0
0.5
1.0
1.5
2.0
1.5
2.0
2.5
3.0
3.5
Carapaxlänge (mm)

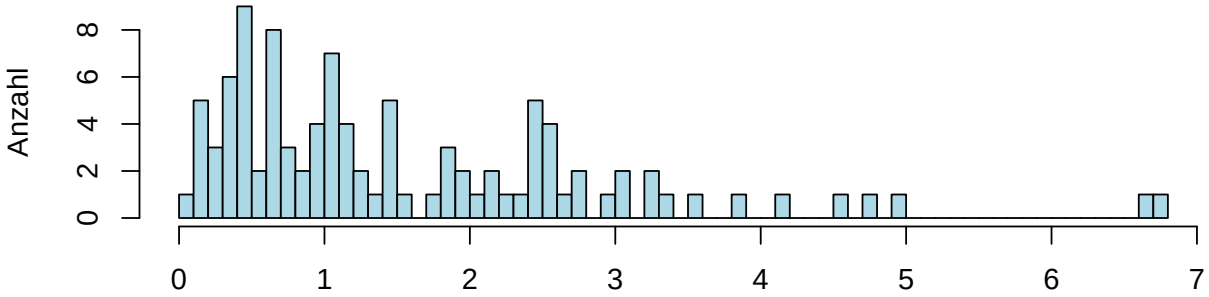
Anzahl vs. Dichte

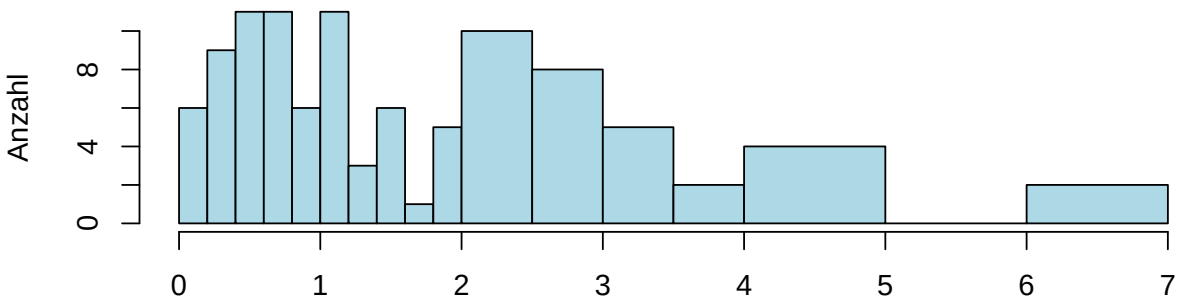


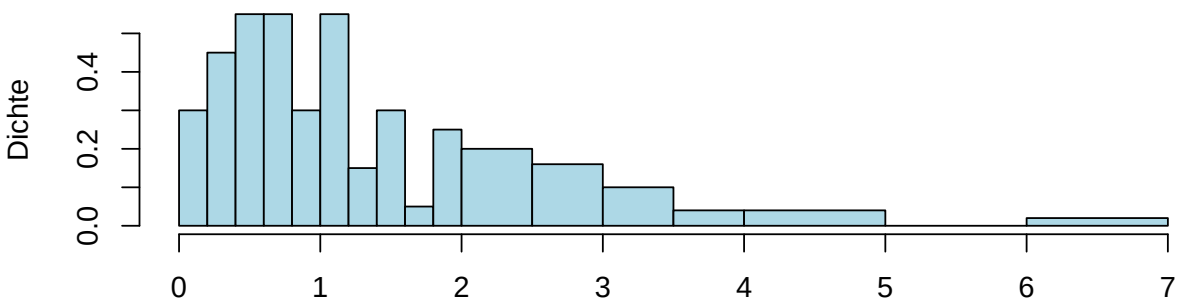
Anzahl vs. Dichte



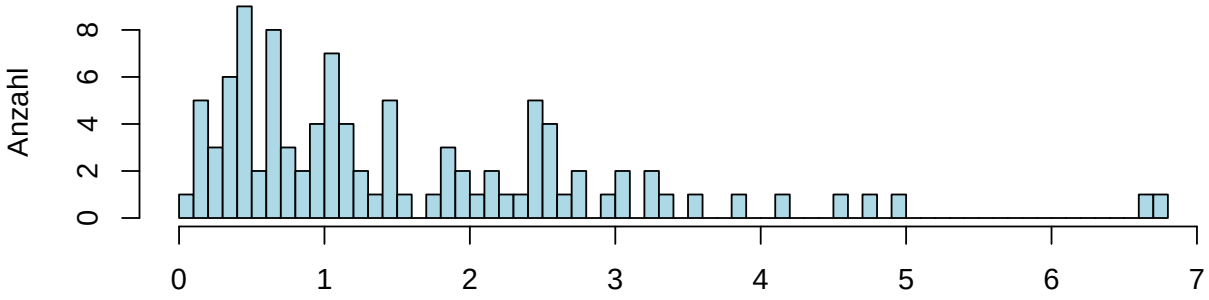
Anzahl vs. Dichte

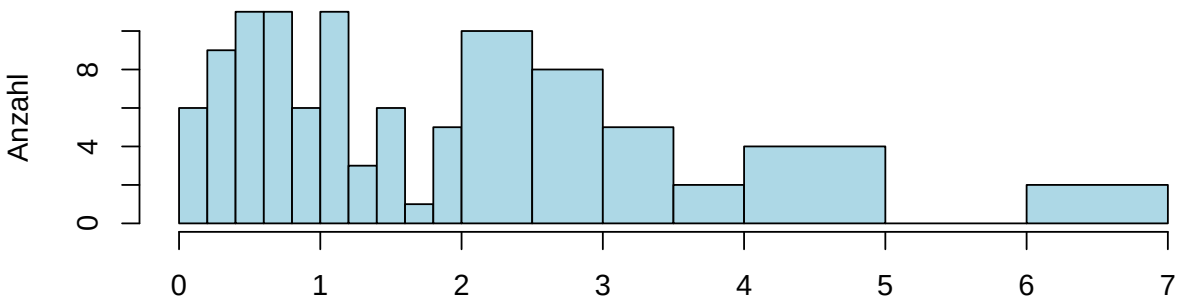


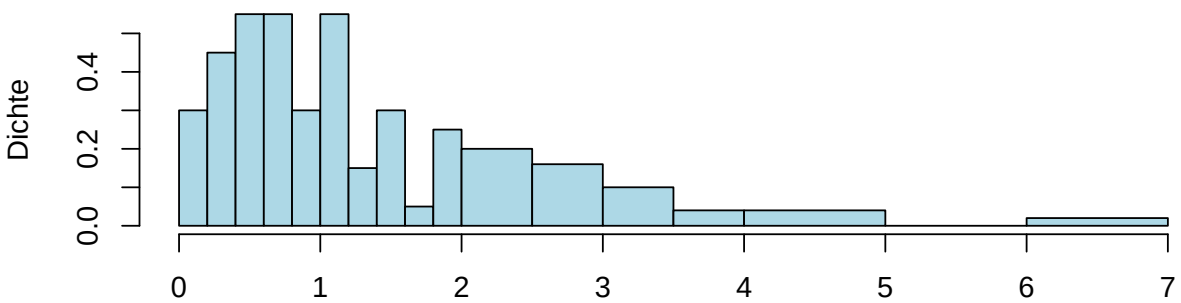




Anzahl vs. Dichte

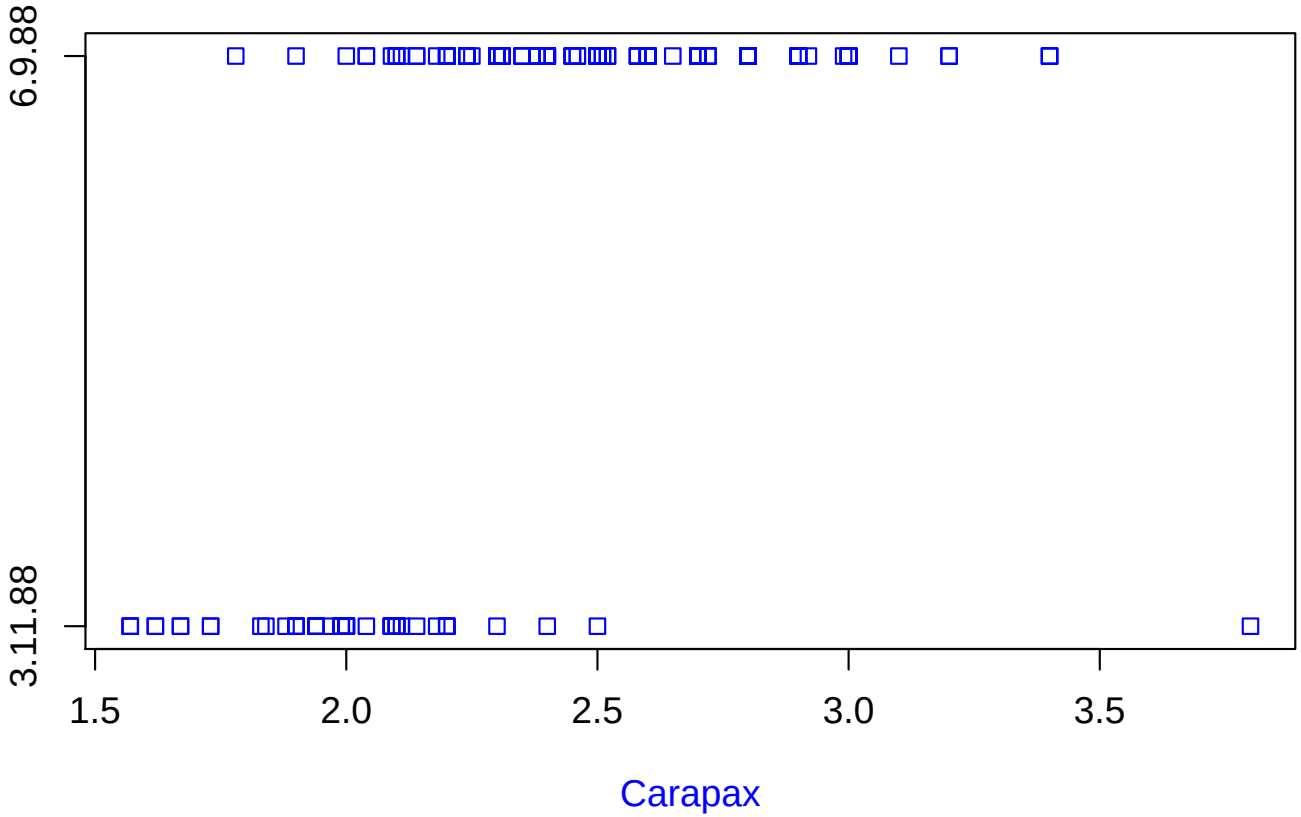


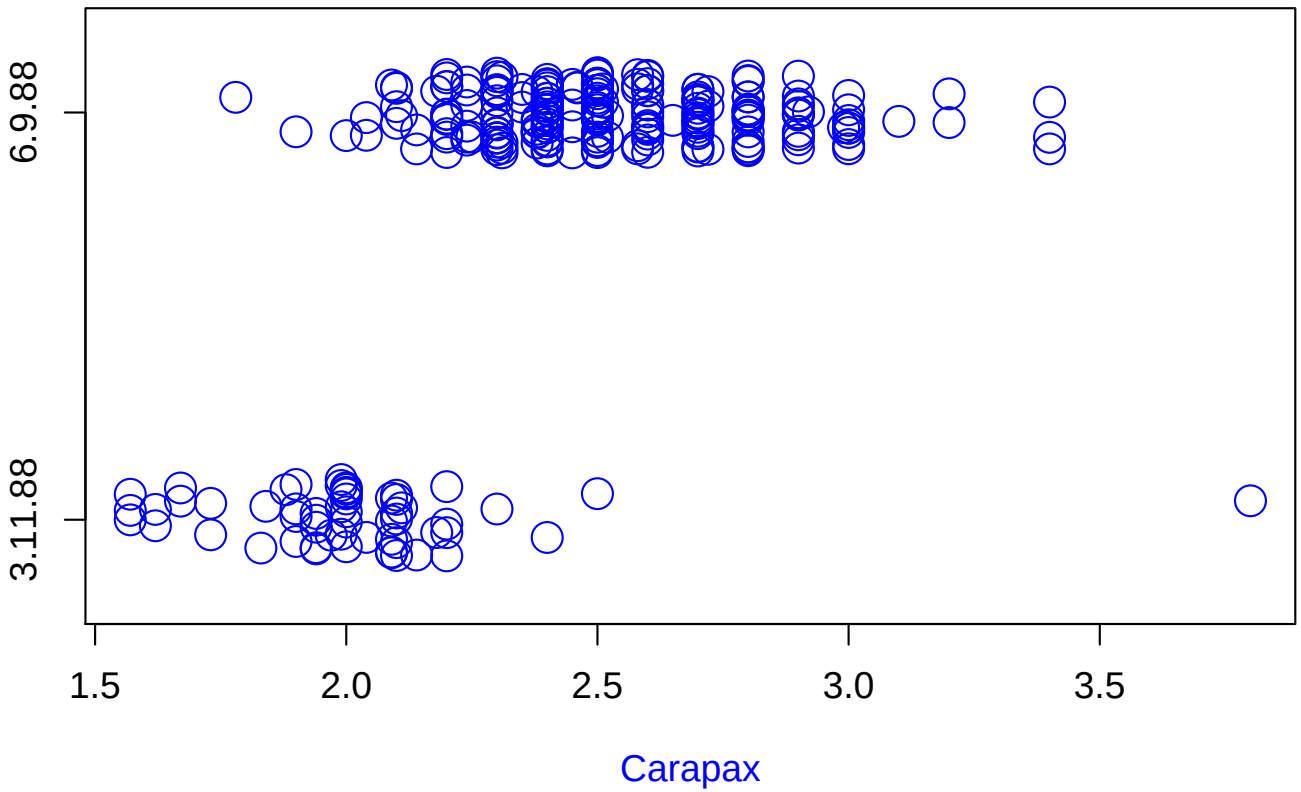


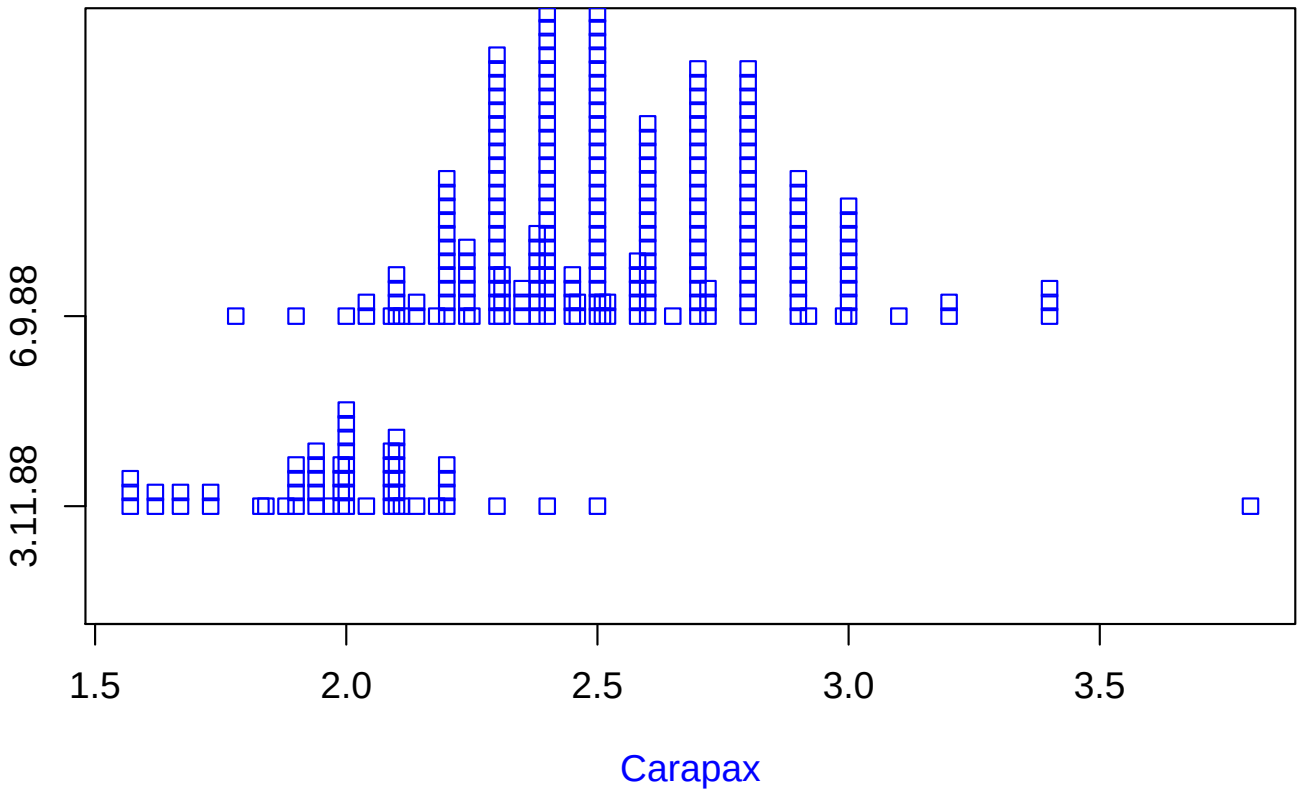


Also: Bei Histogrammen mit ungleichmäßiger Unterteilung immer Dichten verwenden!

Inhalt


6.9.88
3.11.88
1.5
2.0
2.5
3.0
3.5
Carapax


6.9.88
3.11.88
1.5
2.0
2.5
3.0
3.5
Carapax


6.9.88
3.11.88
1.5
2.0
2.5
3.0
3.5
Carapax

Histogramme und Dichtepolygone geben ein ausführliches Bild eines Datensatzes.

Histogramme und Dichtepolygone
geben
ein ausführliches Bild
eines Datensatzes.

Manchmal zu ausführlich.

Inhalt

Zu viel Information erschwert den Überblick

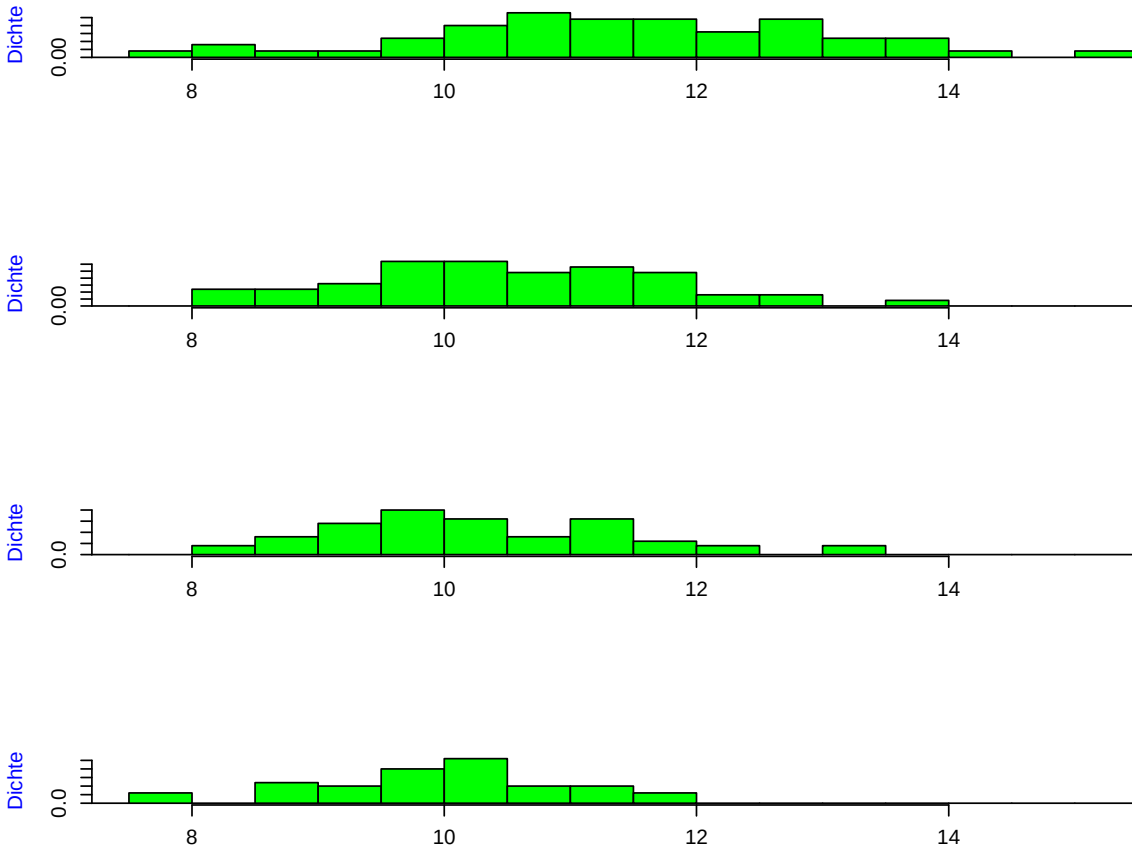
Baum Baum Baum Baum Baum Baum Baum Baum Baum Baum Baum Baum Baum

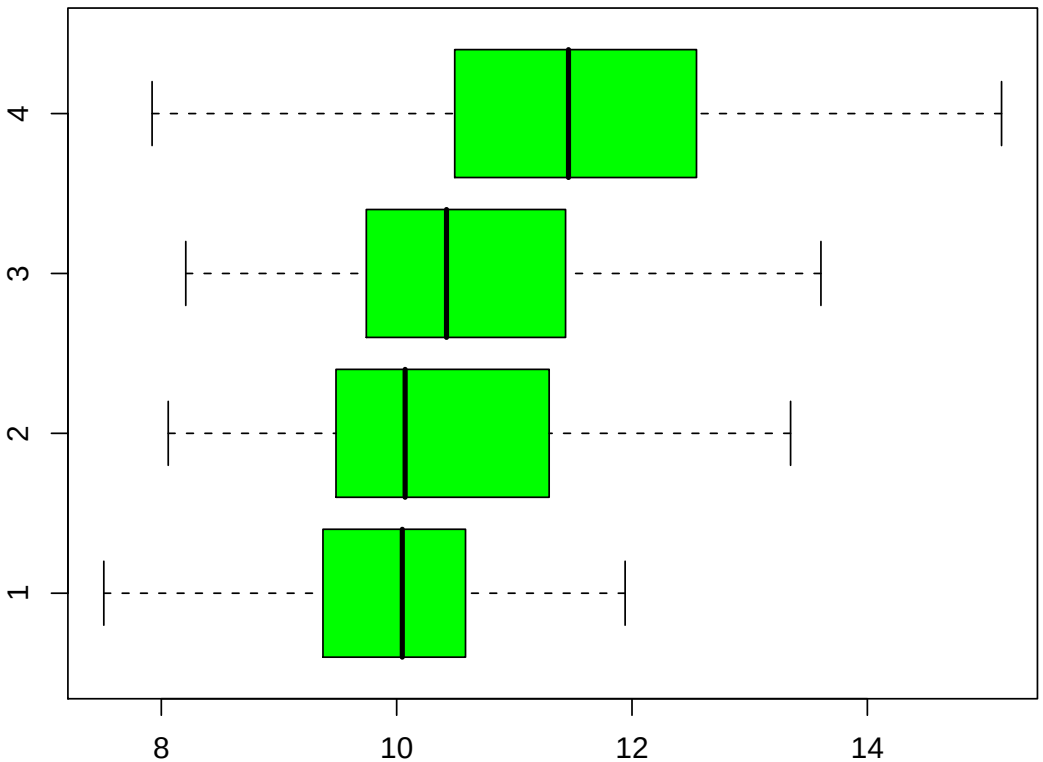
Zu viel Information erschwert den Überblick

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

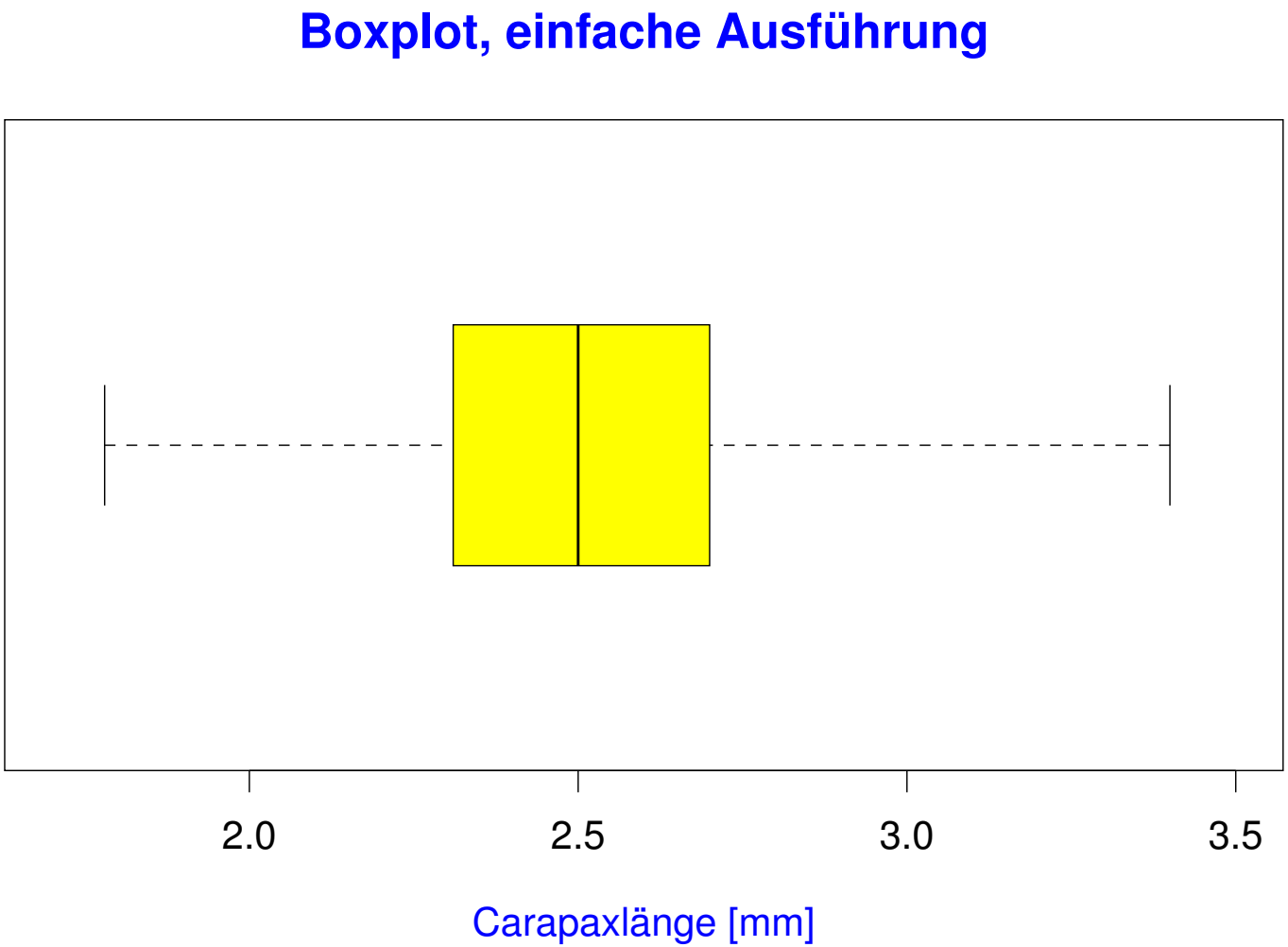
Wald?

Beispiel:
Vergleich von mehreren Gruppen

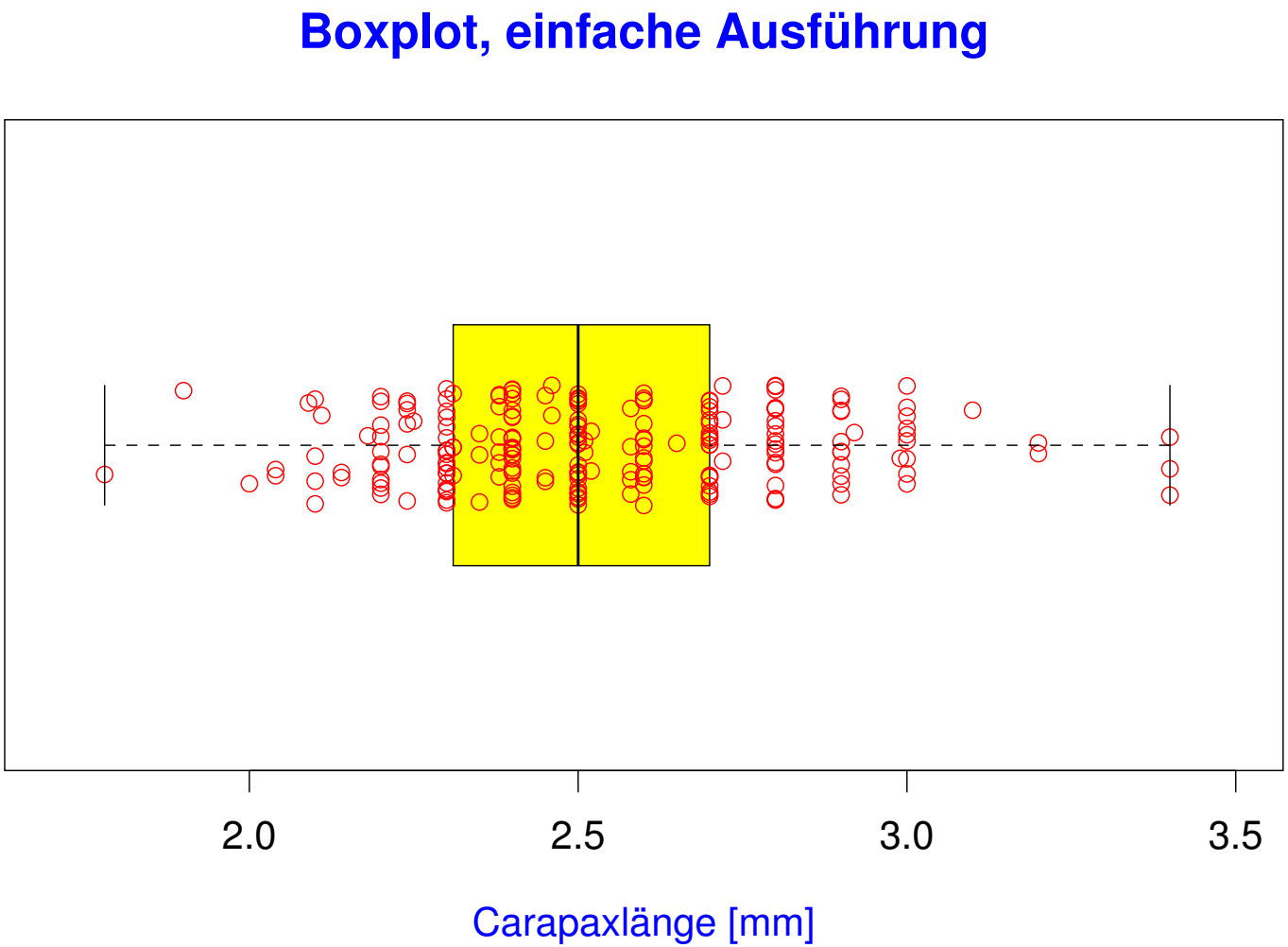

Dichte
0.00
8
10
12
14
Dichte
0.00
8
10
12
14
Dichte
0.0
8
10
12
14
Dichte
0.0
8
10
12
14


4
3
2
1
8
10
12
14

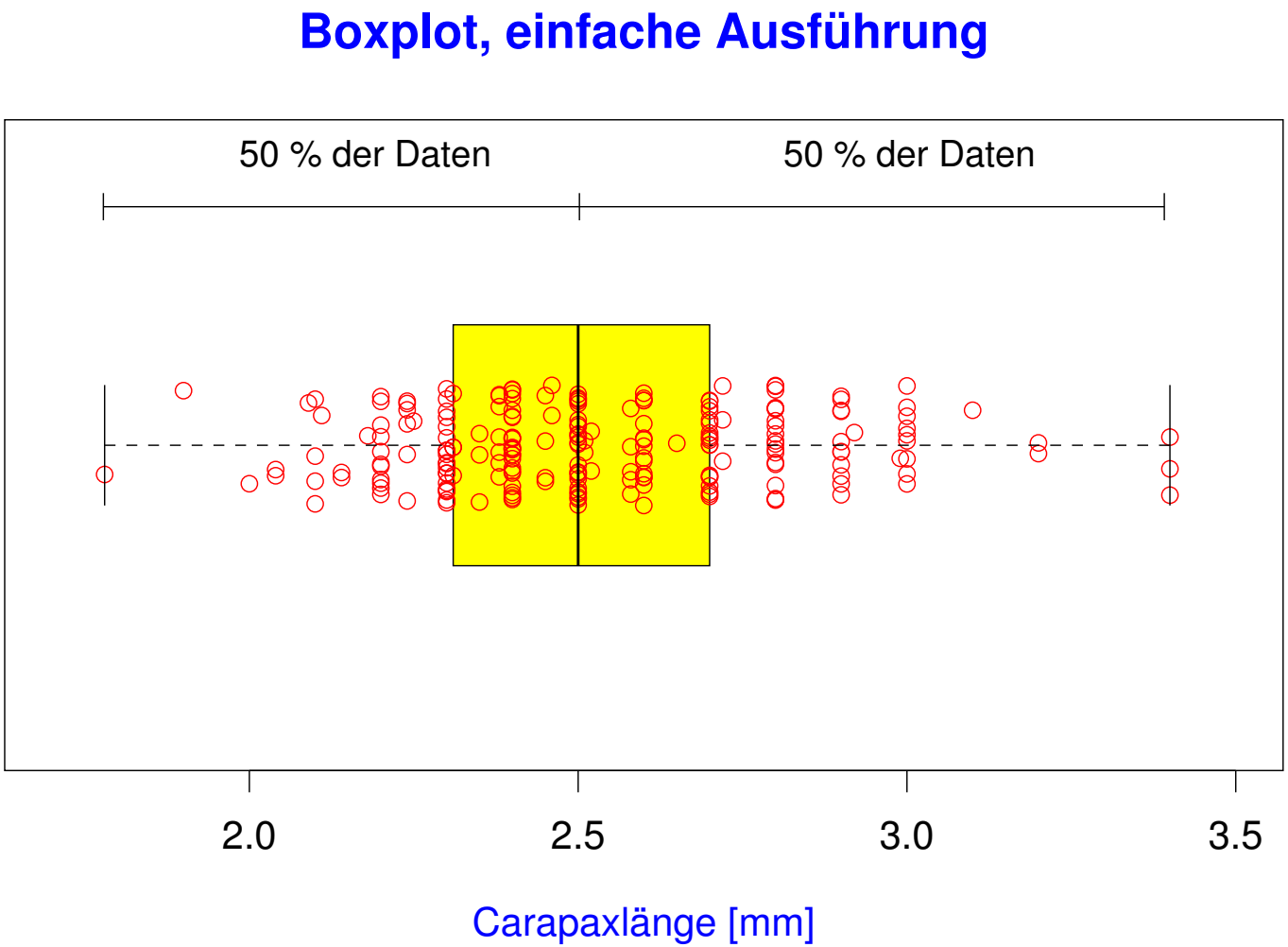
Der Boxplot



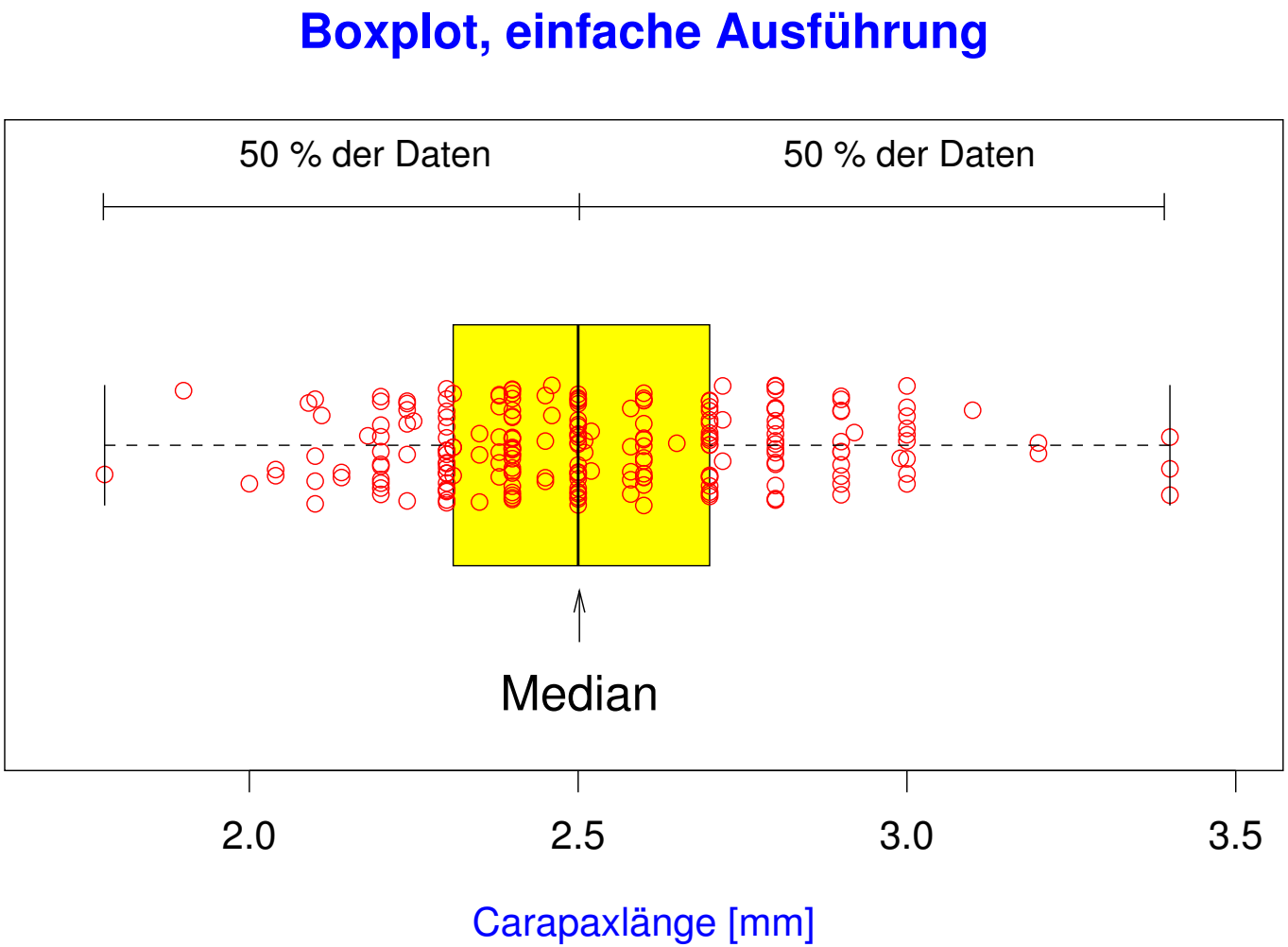
Der Boxplot



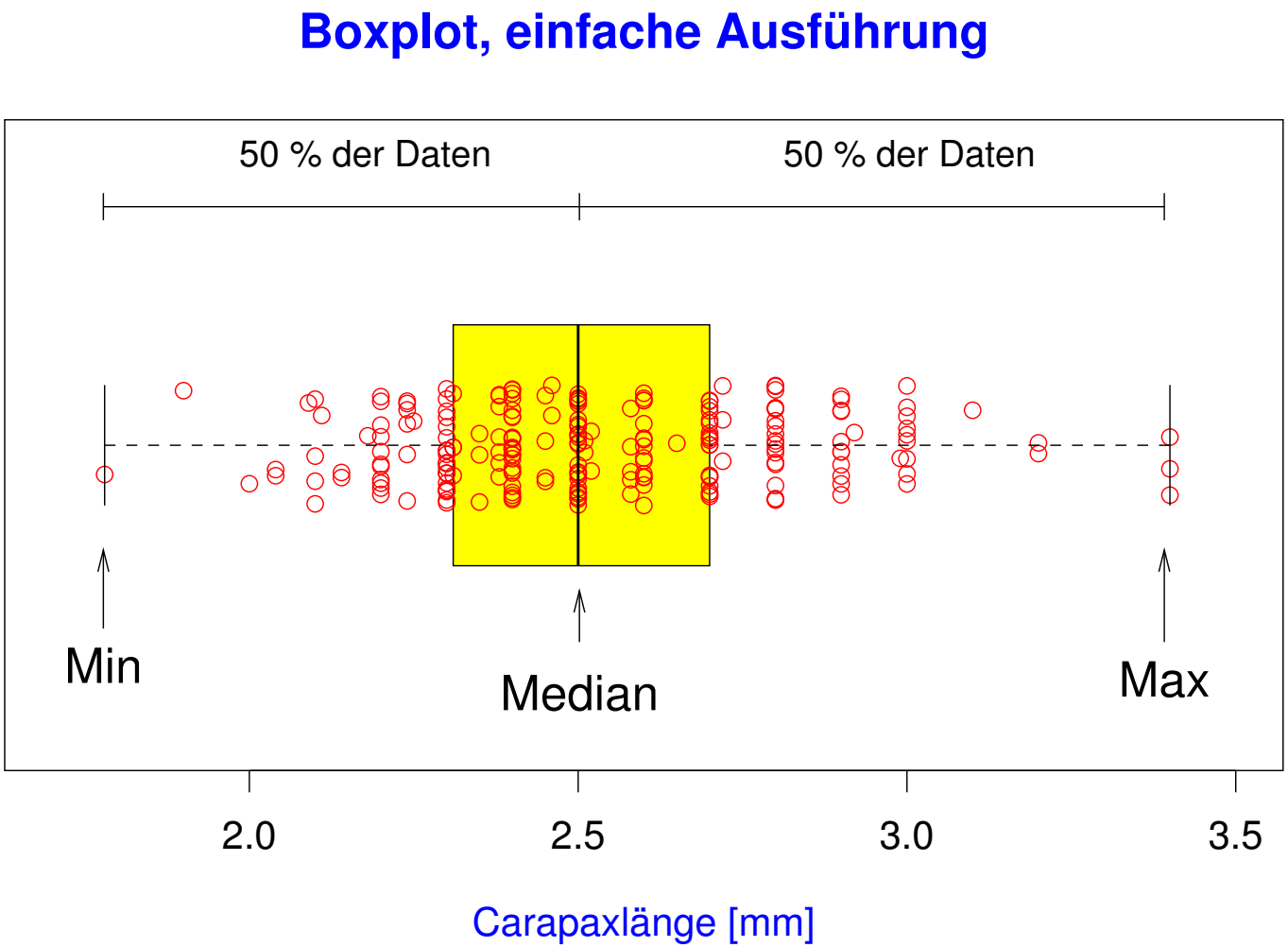
Der Boxplot


Boxplot, einfache Ausführung
50 % der Daten
50 % der Daten
2.0
2.5
3.0
3.5
Carapaxlänge [mm]

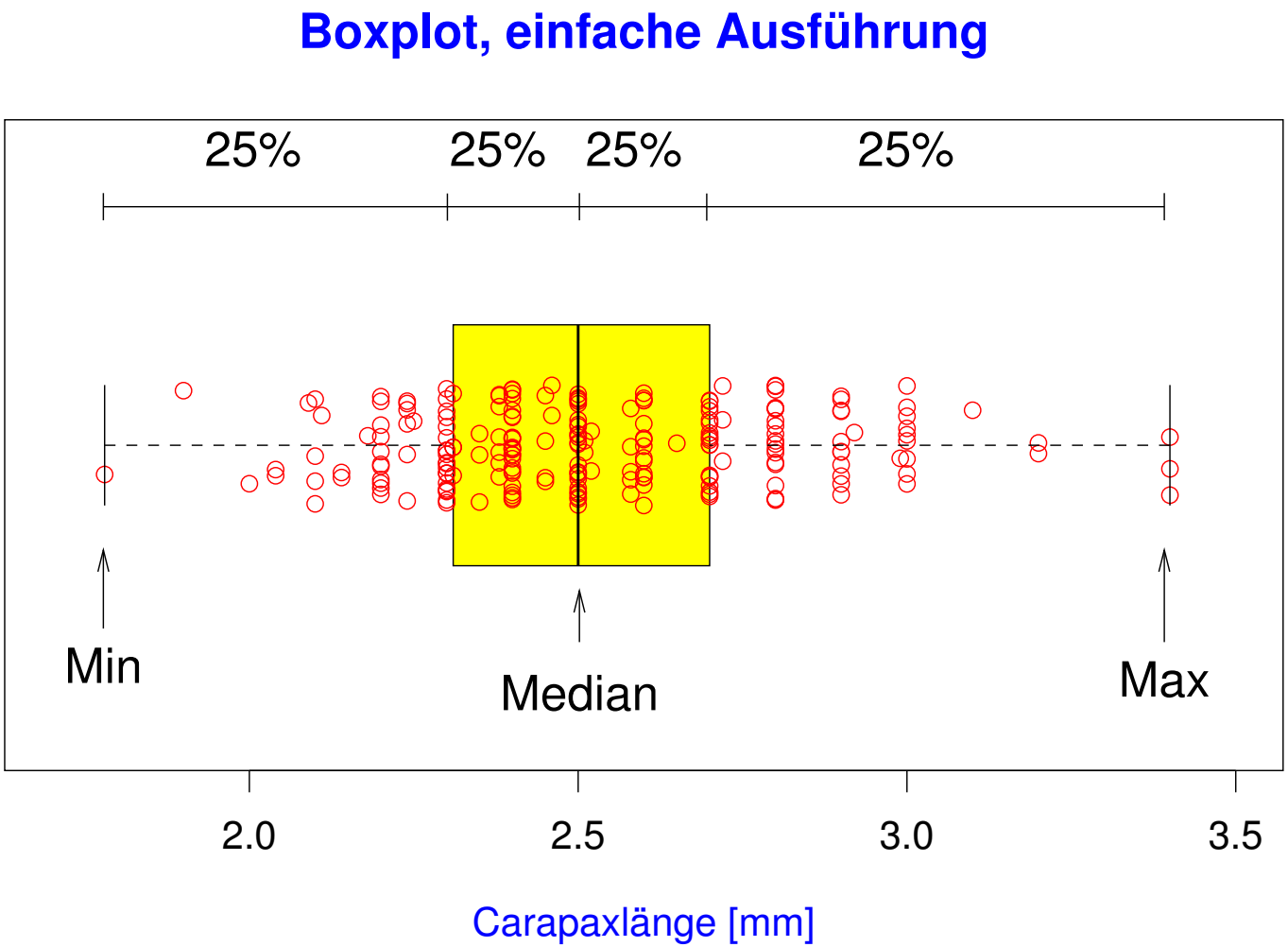
Der Boxplot


Boxplot, einfache Ausführung
50 % der Daten
50 % der Daten
Median
2.0
2.5
3.0
3.5
Carapaxlänge [mm]

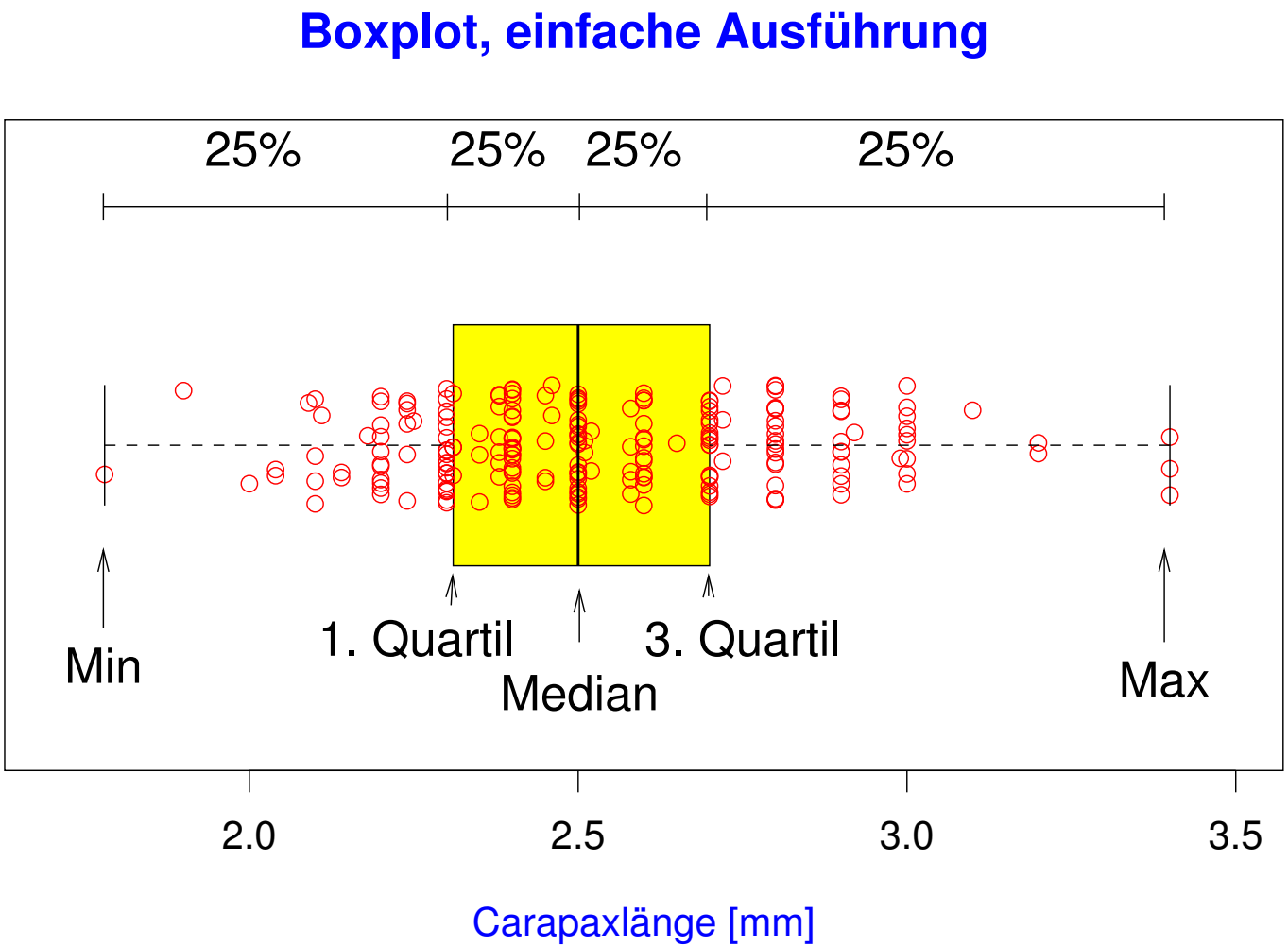
Der Boxplot


Boxplot, einfache Ausführung
50 % der Daten
50 % der Daten
Min
Median
Max
2.0
2.5
3.0
3.5
Carapaxlänge [mm]

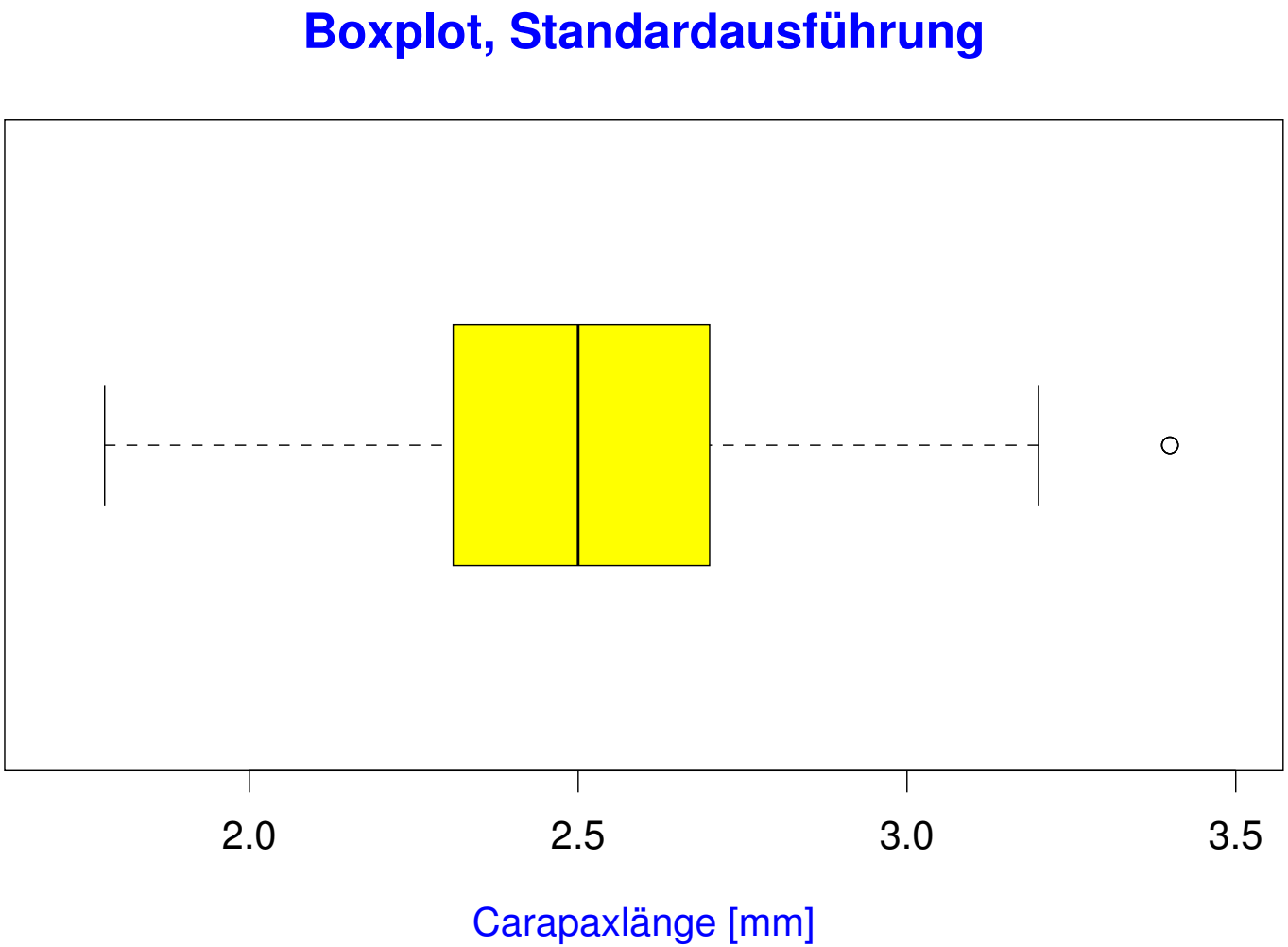
Der Boxplot


Boxplot, einfache Ausführung
25%
25%
25%
25%
Min
Median
Max
2.0
2.5
3.0
3.5
Carapaxlänge [mm]

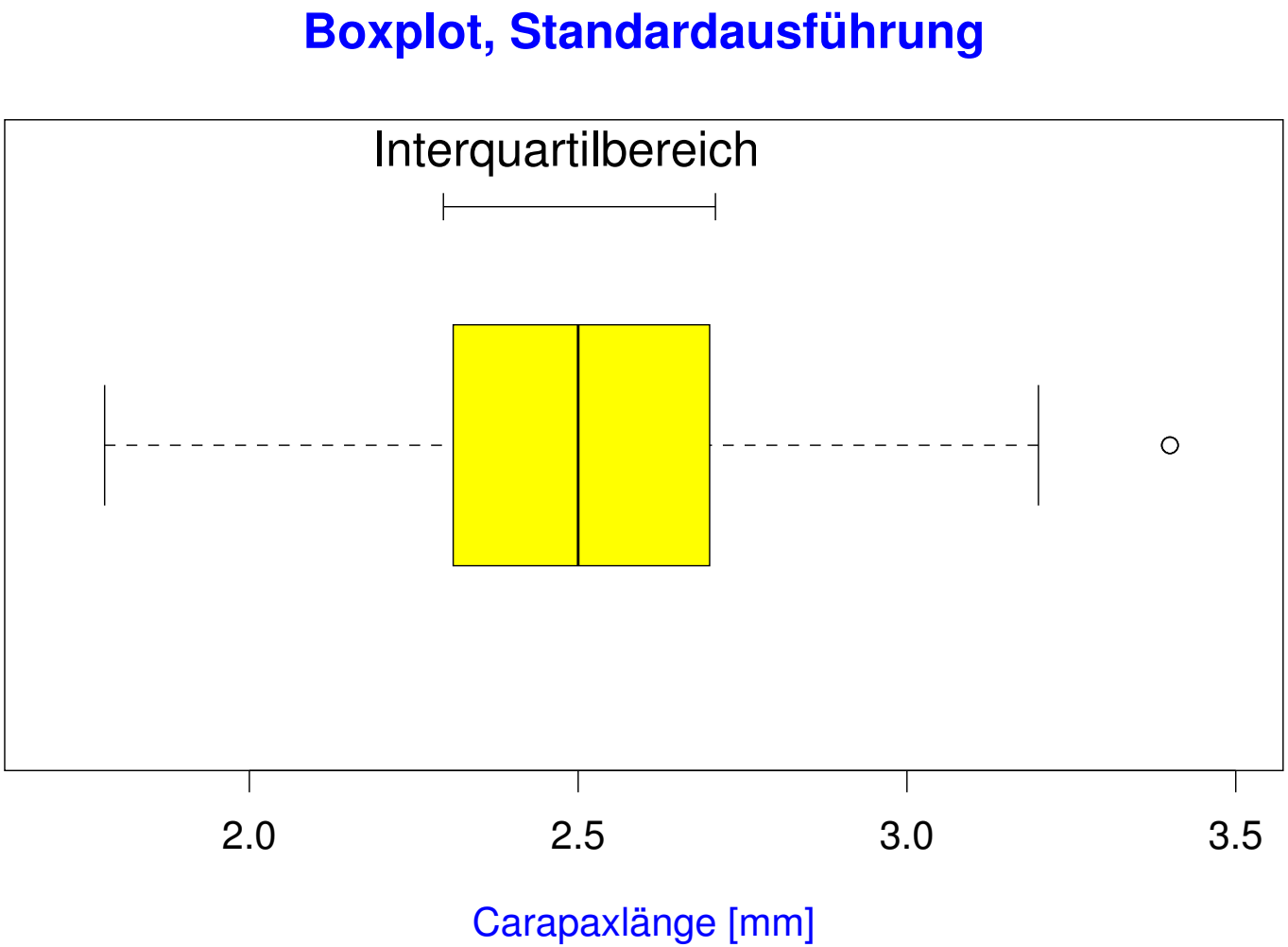
Der Boxplot


Boxplot, einfache Ausführung
25%
25%
25%
25%
Min
1. Quartil
Median
3. Quartil
Max
2.0
2.5
3.0
3.5
Carapaxlänge [mm]

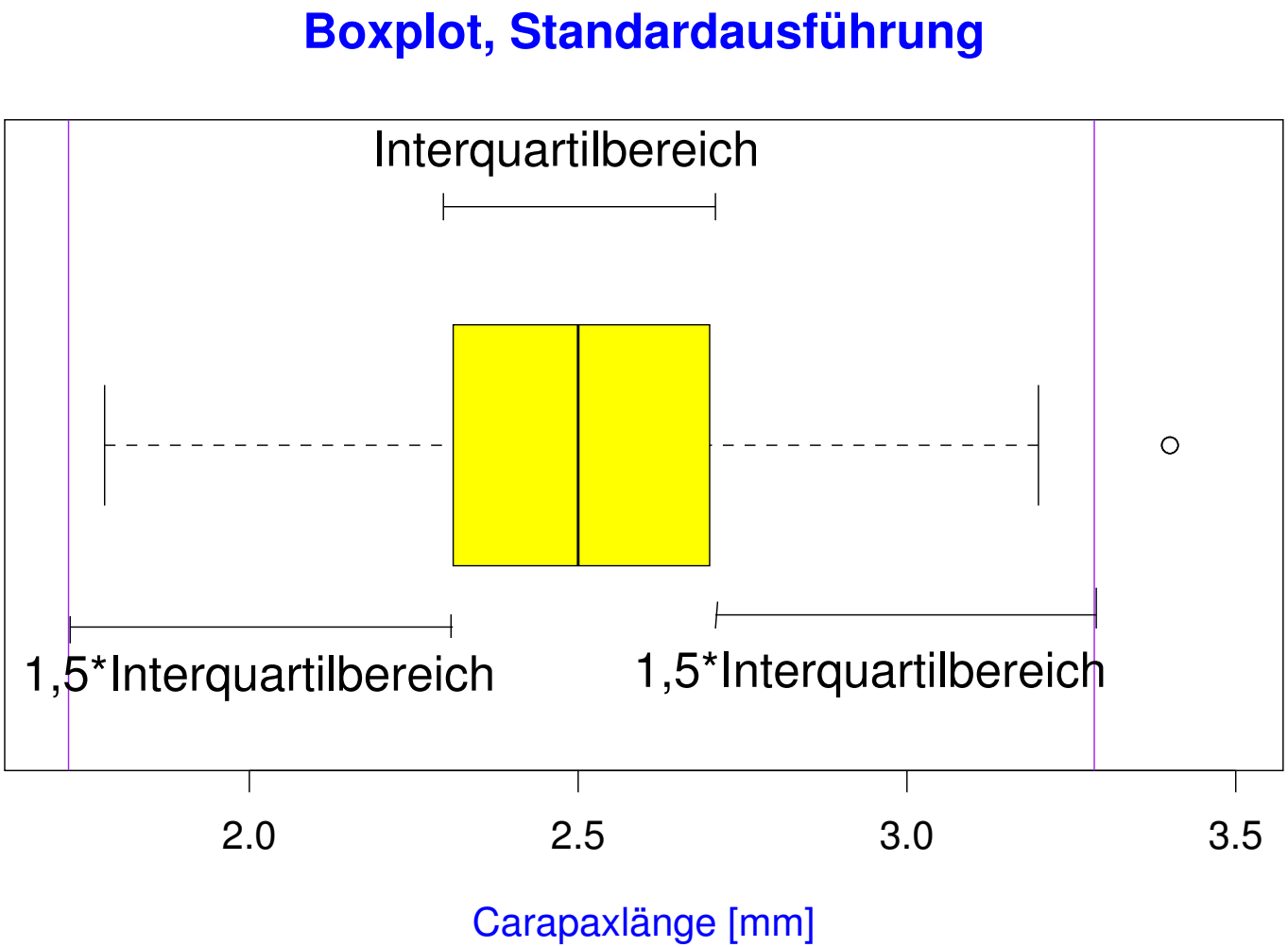
Der Boxplot



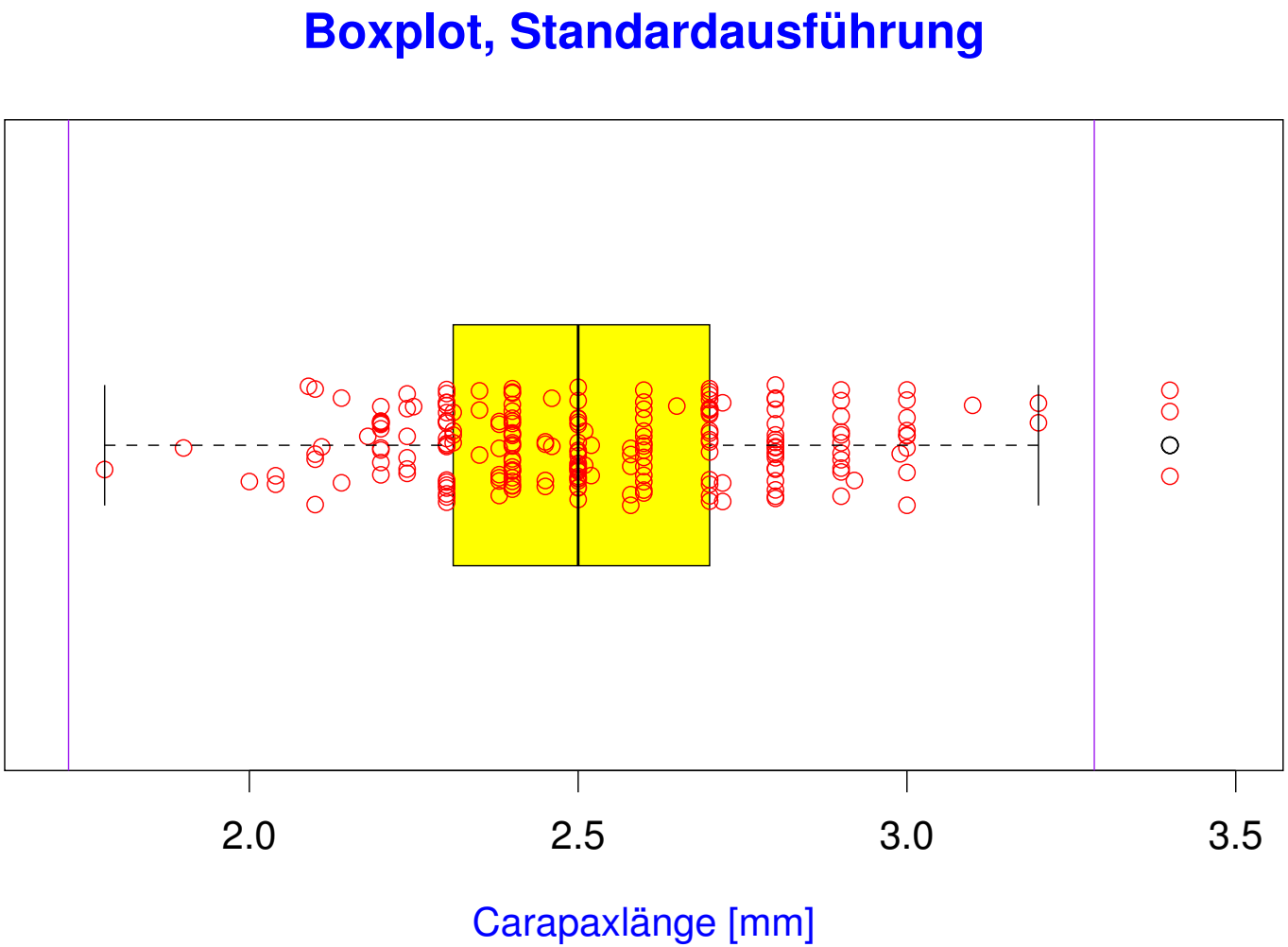
Der Boxplot


Boxplot, Standardausführung
Interquartilbereich
2.0
2.5
3.0
3.5
Carapaxlänge [mm]

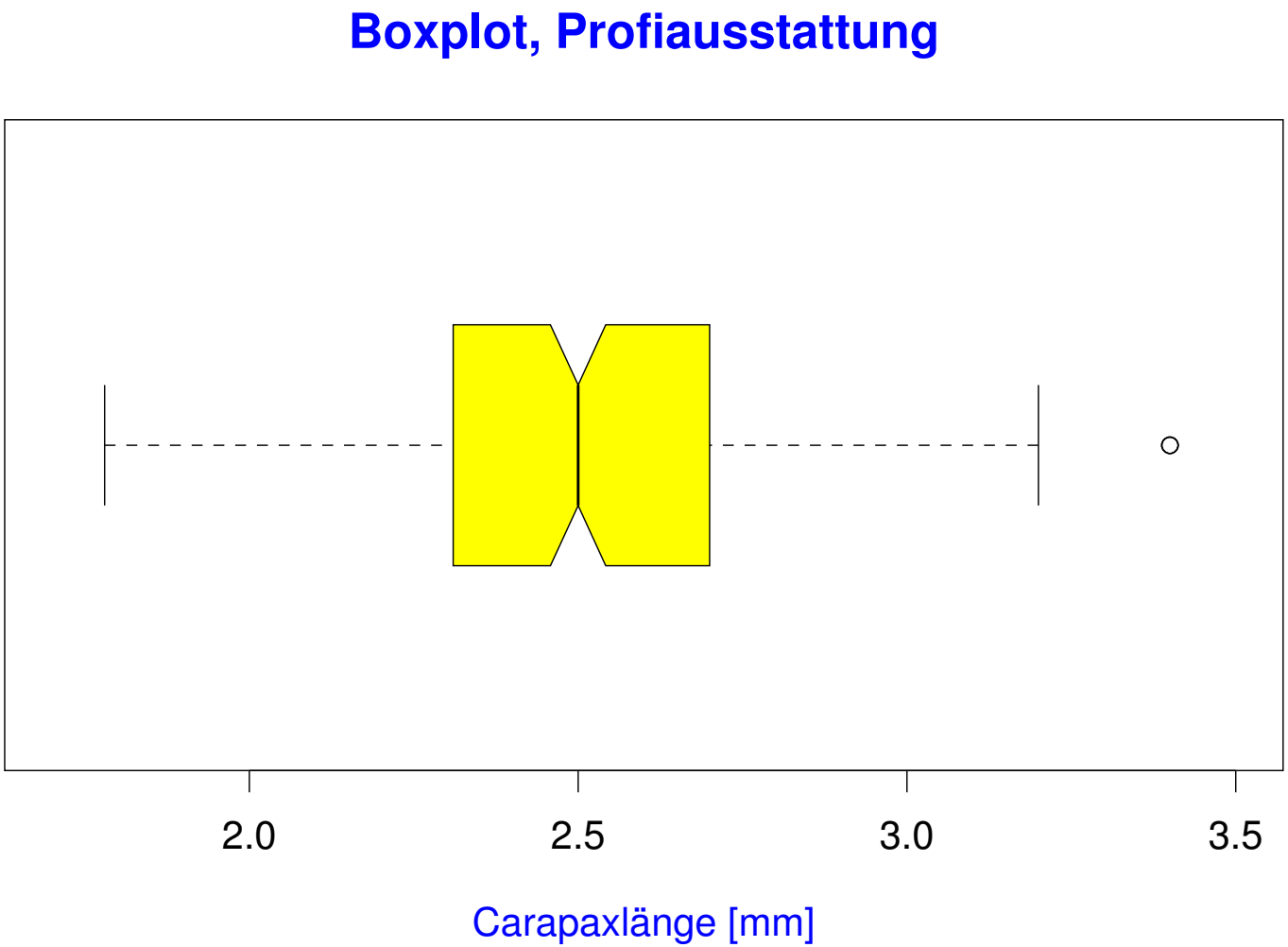
Der Boxplot


Boxplot, Standardausführung
Interquartilbereich
1,5*Interquartilbereich
1,5*Interquartilbereich
2.0
2.5
3.0
3.5
Carapaxlänge [mm]

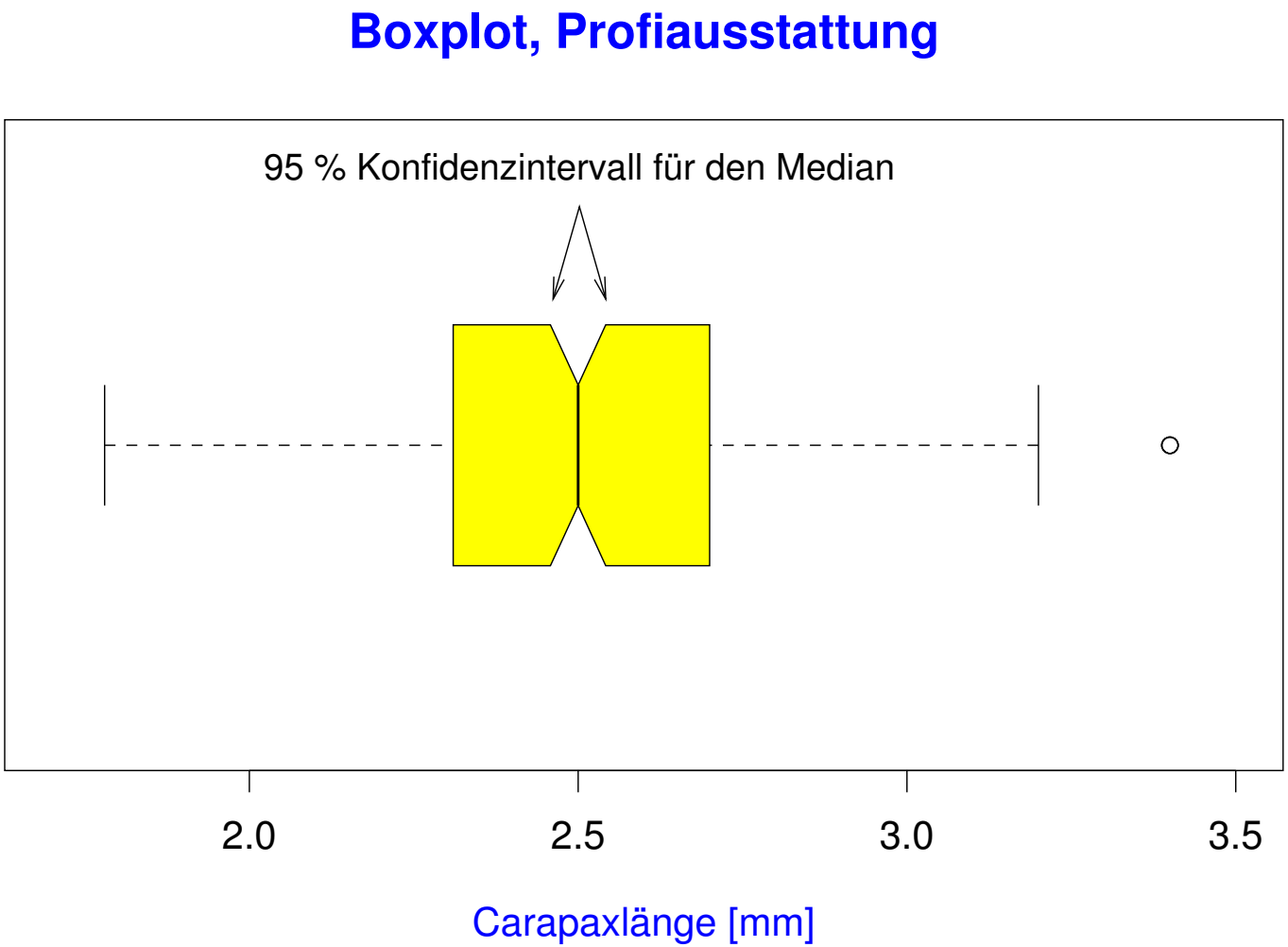
Der Boxplot



Der Boxplot



Der Boxplot


Boxplot, Profiausstattung
95 % Konfidenzintervall für den Median
2.0
2.5
3.0
3.5
Carapaxlänge [mm]

Fazit

1. Histogramme erlauben einen detailierten Blick auf die Daten

Fazit

1. Histogramme erlauben einen detailierten Blick auf die Daten
2. Dichtepolygone erlauben Vergleiche zwischen vielen Verteilungen

Fazit

1. Histogramme erlauben einen detailierten Blick auf die Daten
2. Dichtepolygone erlauben Vergleiche zwischen vielen Verteilungen
3. Boxplot können große Datenmengen vereinfacht zusammenfassen

Fazit

1. Histogramme erlauben einen detailierten 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

Fazit

1. Histogramme erlauben einen detailierten 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

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) α-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 $x_1, x_2, \ldots, x_n$ gegeben, $\alpha \in (0, 1)$.

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

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

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

Seien $x_{(1)} \leq x_{(2)} \leq \cdots \leq x_{(n)}$ 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 $\left|\{i : x_i \leq x_{(k)}\}\right| \geq k$, $\left|\{i : x_i \geq x_{(k)}\}\right| \geq n - k + 1$.

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

(Empirische) Quantile, allgemein II

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

(ein) α-Quantil q der n 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) α-Quantil q der n Beobachtungswerte erfüllt

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

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.)
- Die Uneindeutigkeit des α-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. σ)

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. σ)

wobei

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

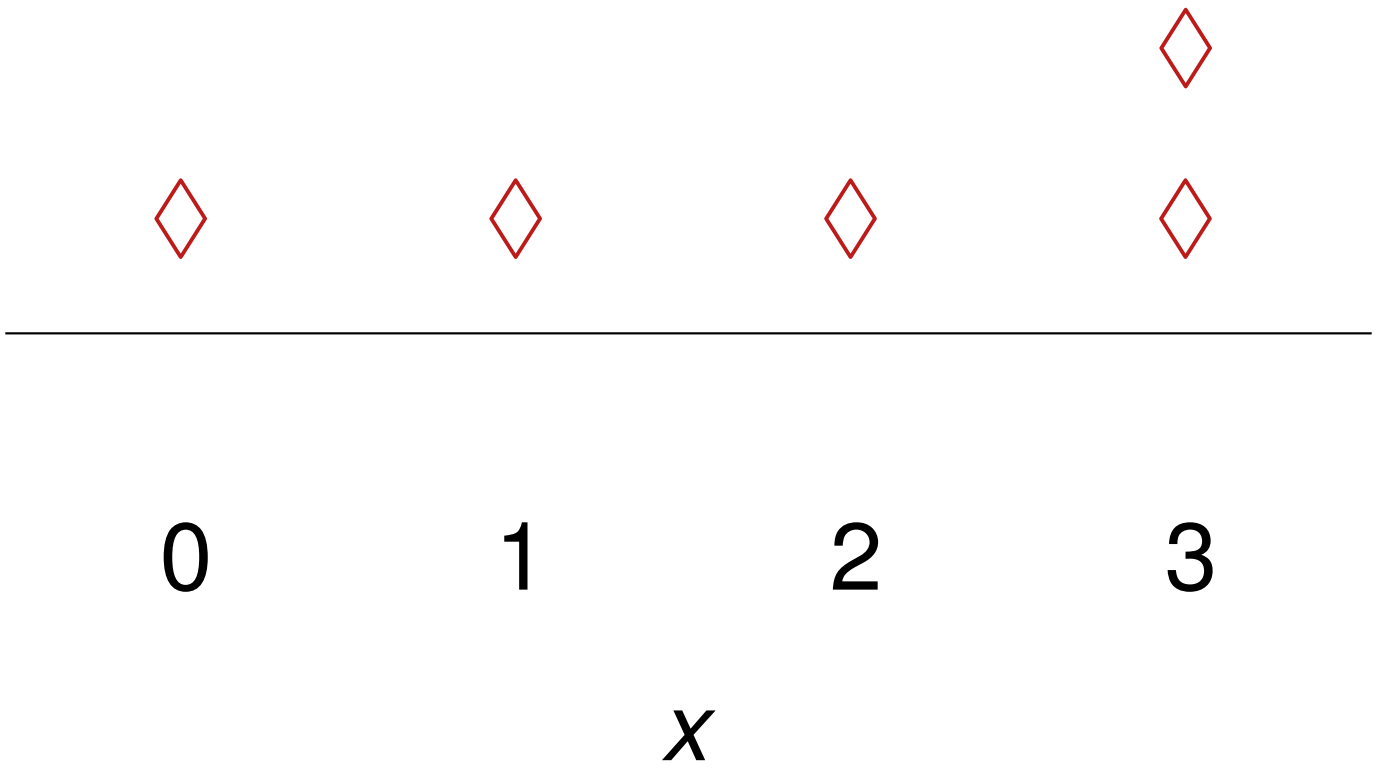
$s^2 = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \overline{x})^2$ die korrigierte Stichproben-Varianz

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

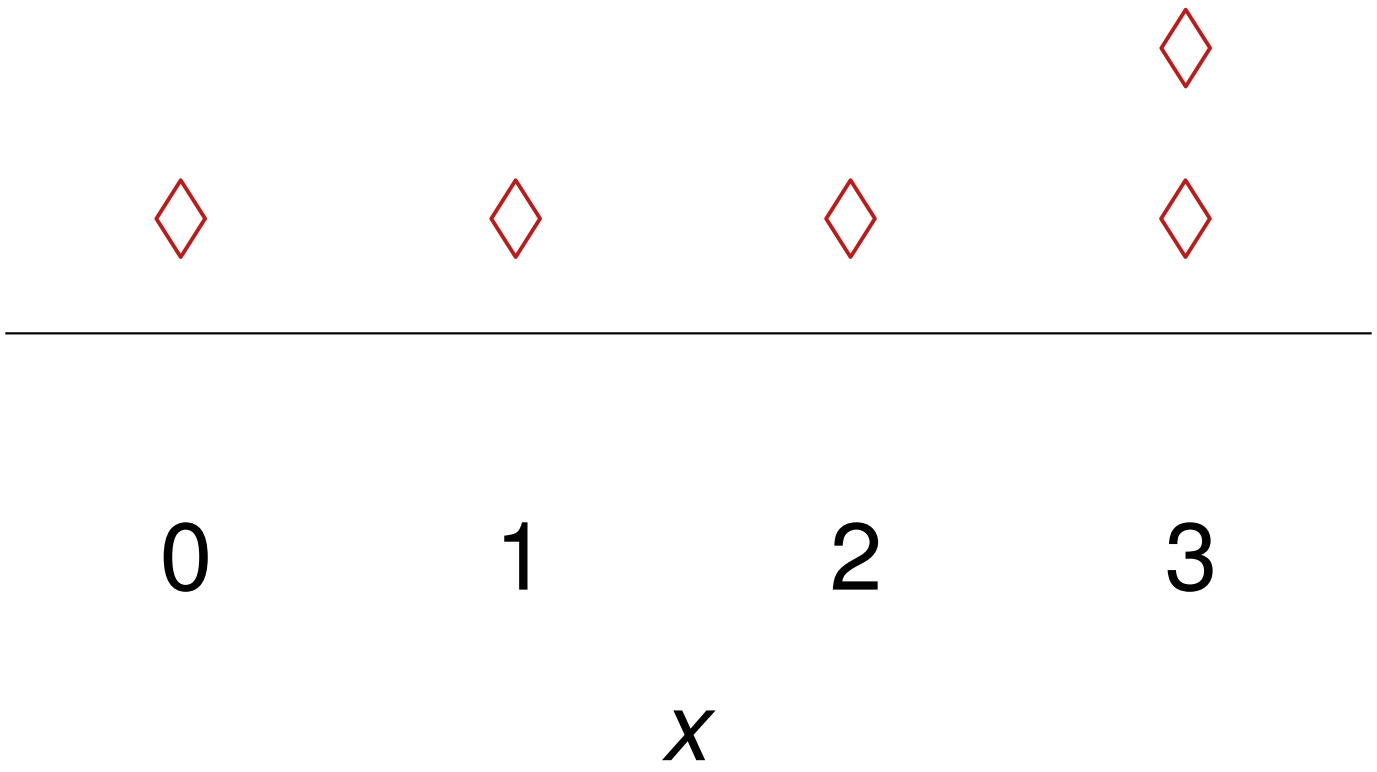
Erinnerung: Geometrische Bedeutung des Mittelwerts

Der Schwerpunkt

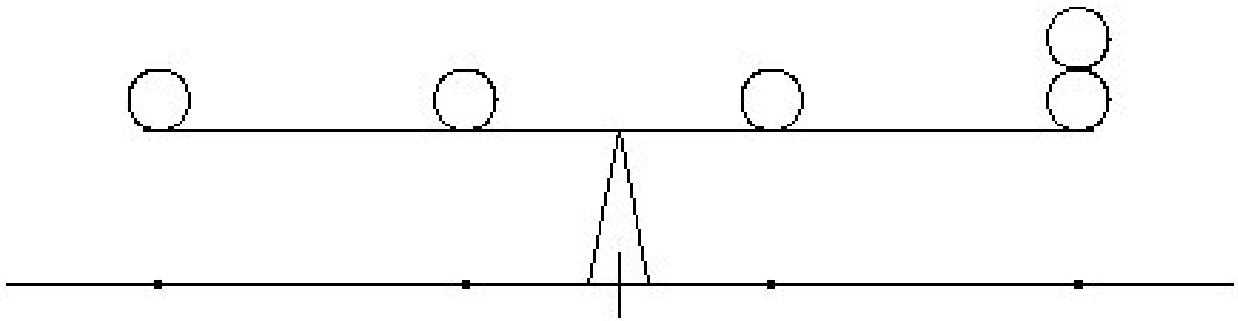
Wir stellen uns die Beobachtungen als gleich schwere Gewichte auf einer Waage vor:



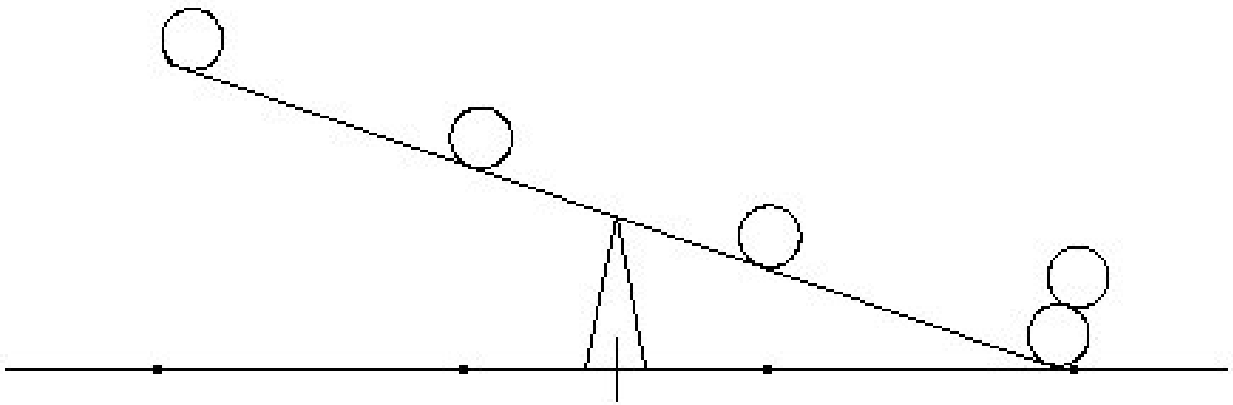
Wo muß der Drehpunkt sein, damit die Waage im Gleichgewicht ist?



$$m = 1{,}5\ ?$$

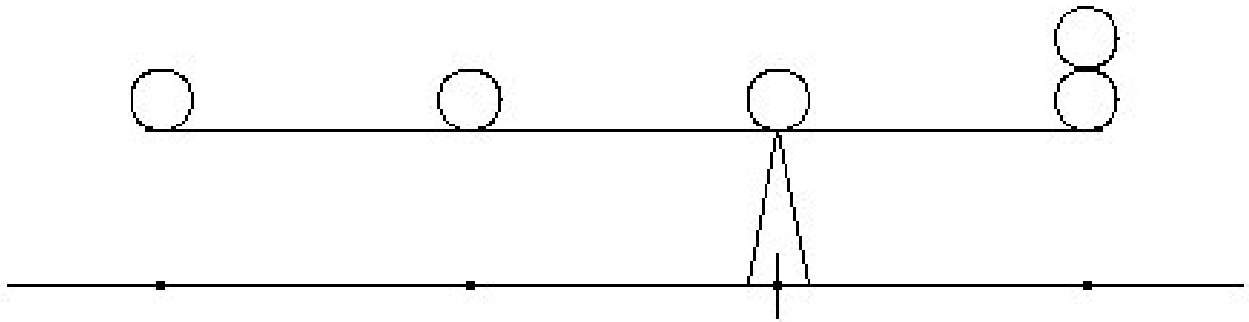


$m = 1{,}5$?

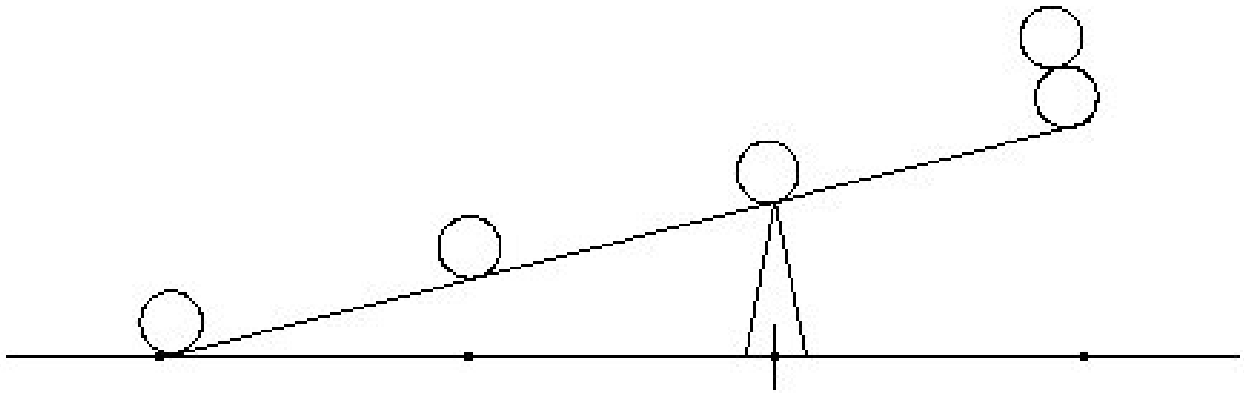


zu klein

$$m = 2\ ?$$

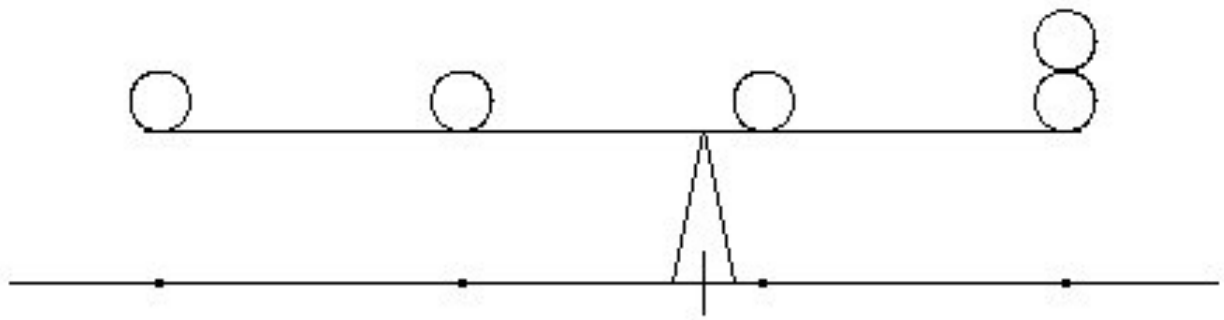


$m = 2$?

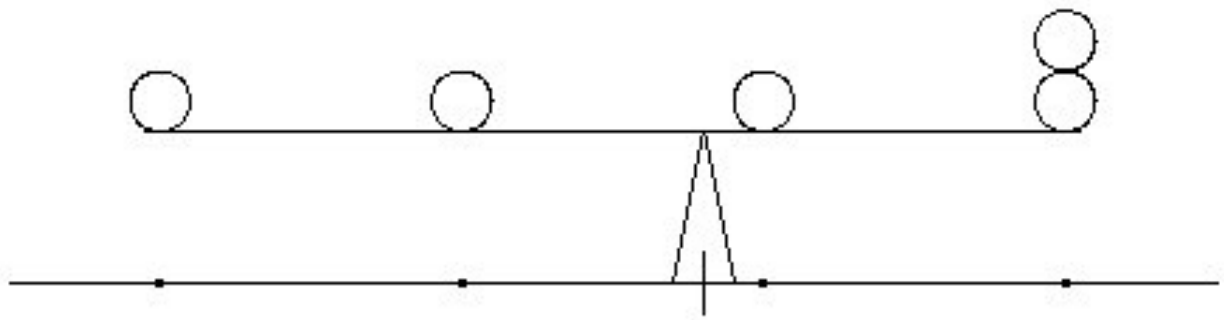


zu groß

$$m = 1{,}8\ ?$$



$m = 1{,}8$?

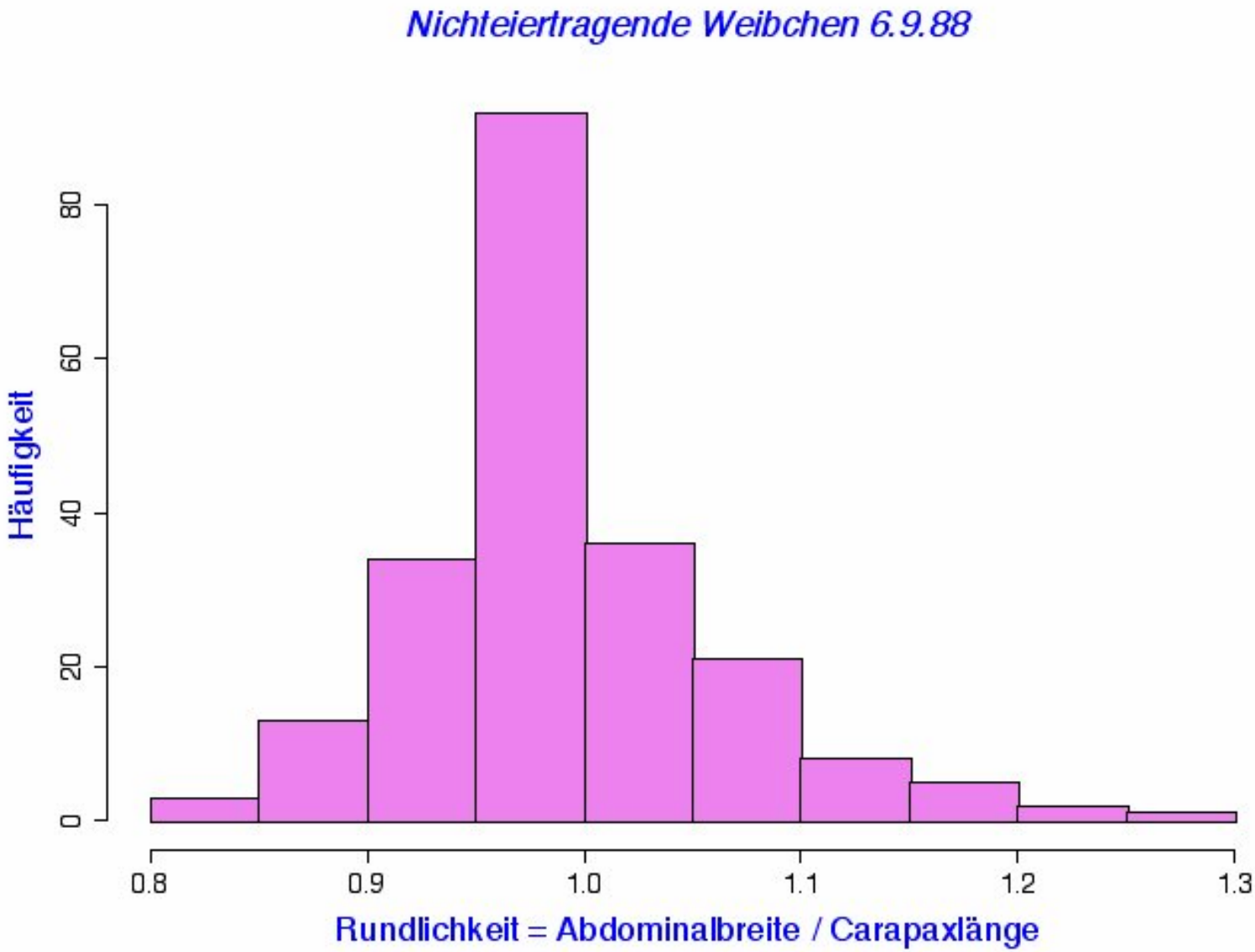


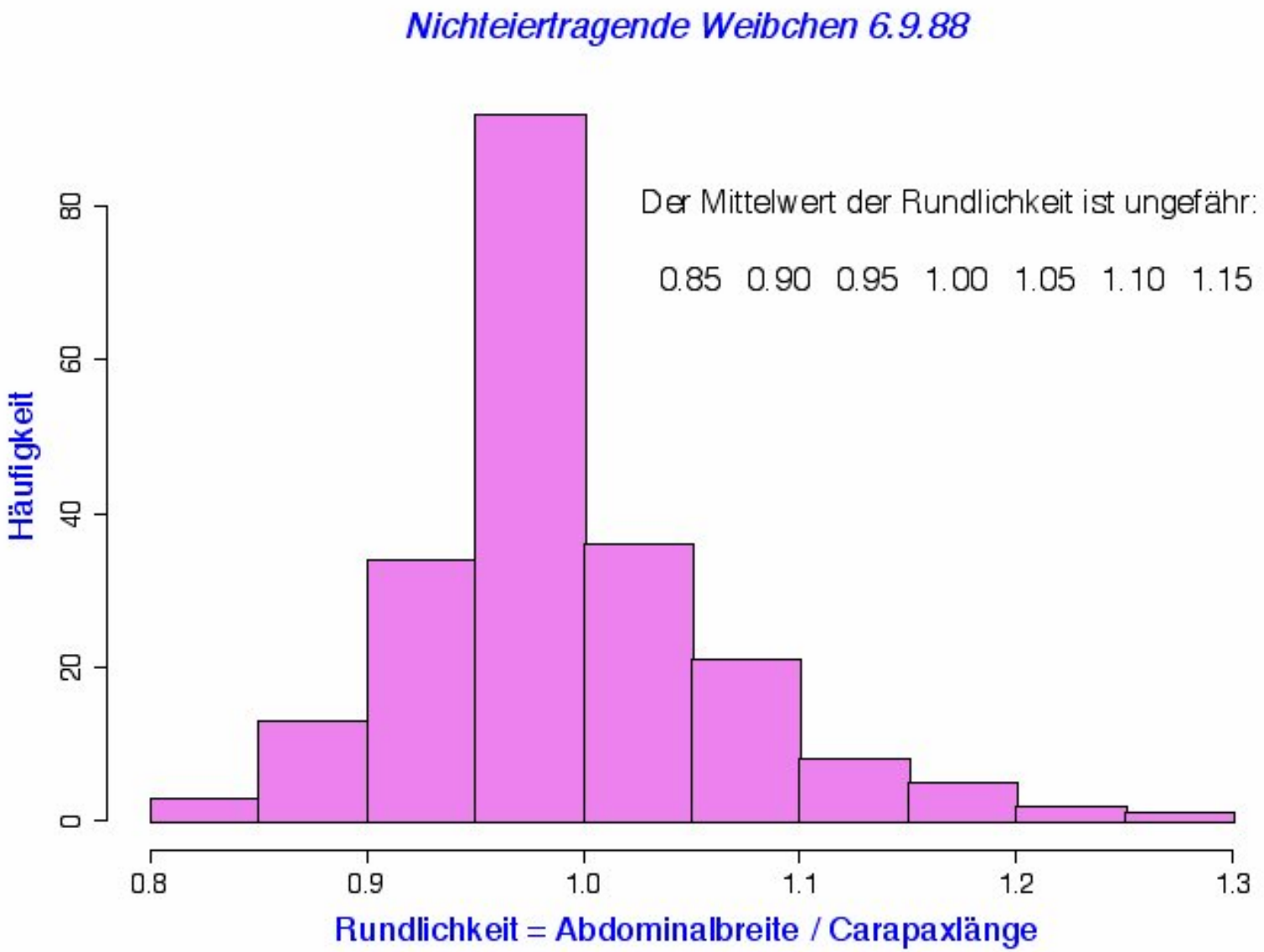
richtig

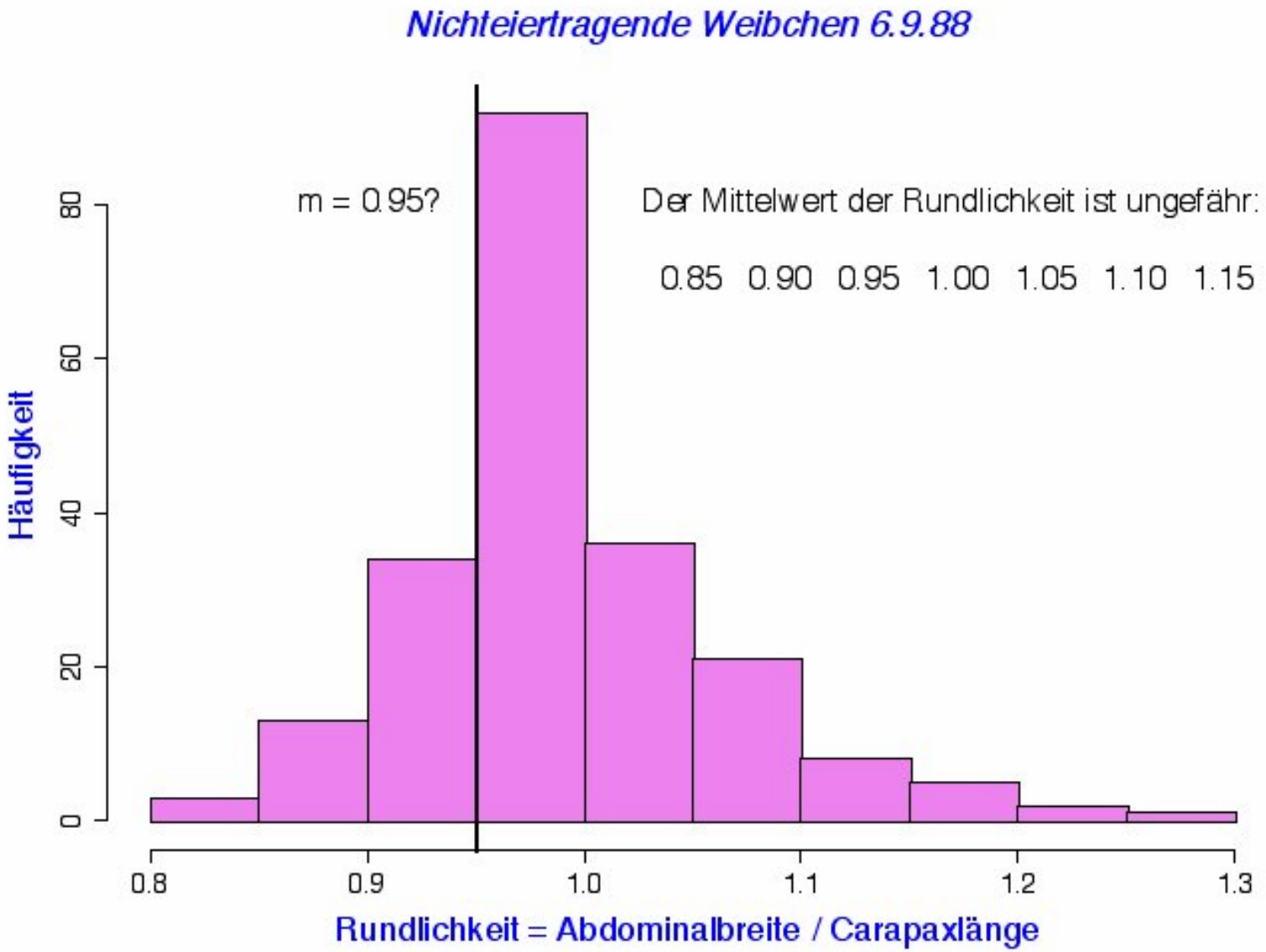
Oft kann man „mit dem bloßen Auge“ anhand eines Histogramms den Mittelwert gut einschätzen.

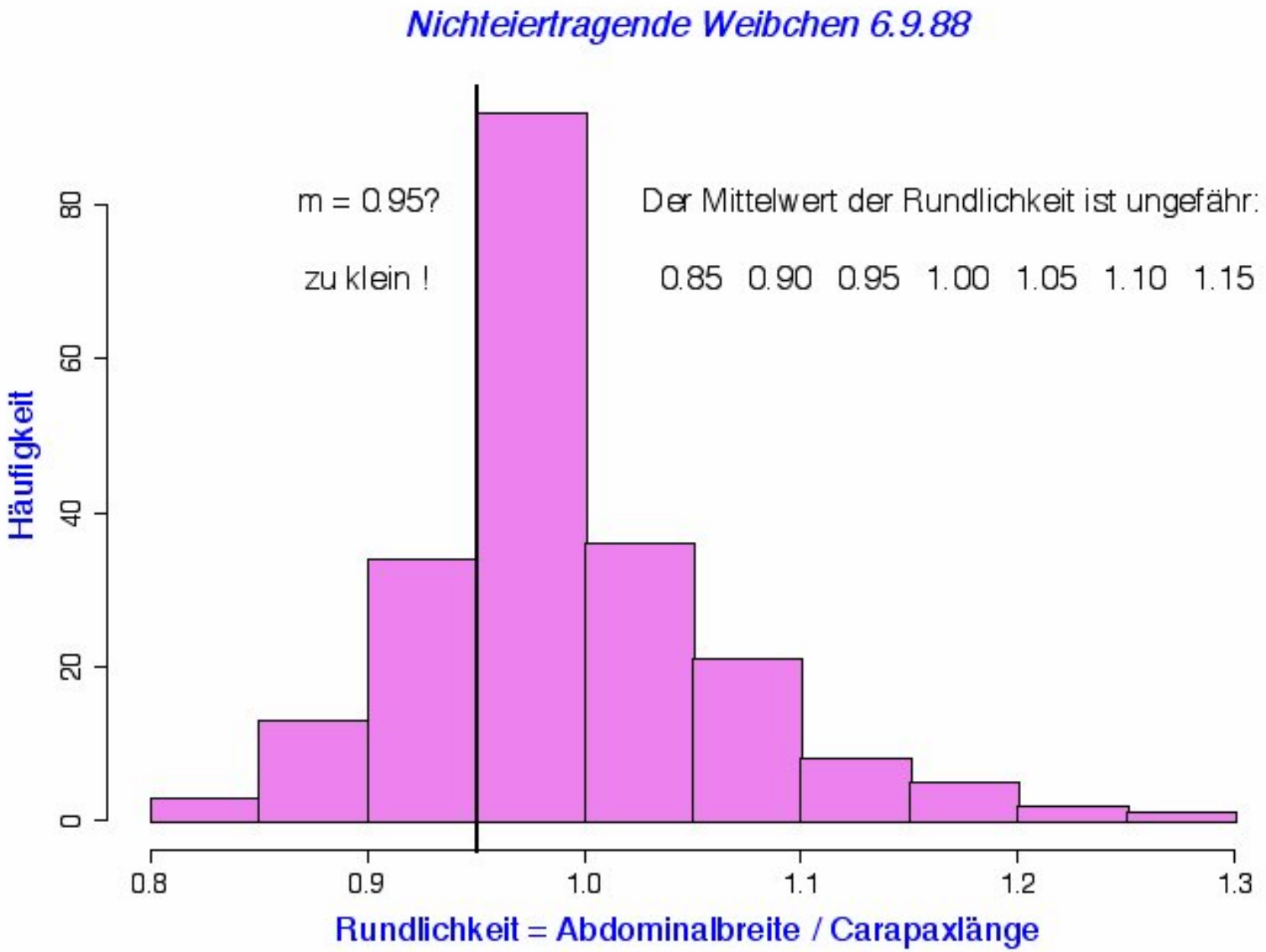
Beispiel: *Galathea intermedia*

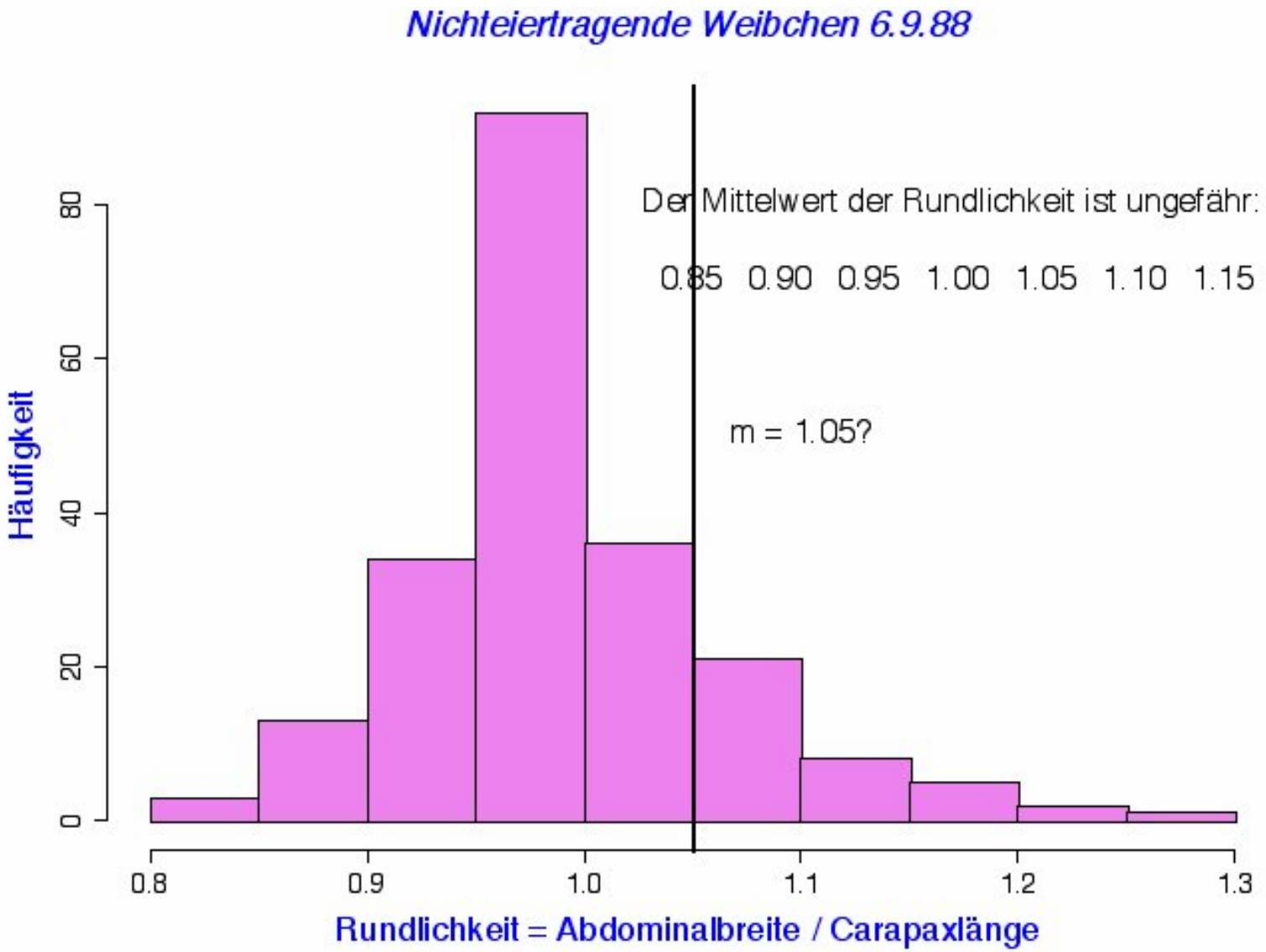
„Rundlichkeit“
:=
Abdominalbreite / Carapaxlänge

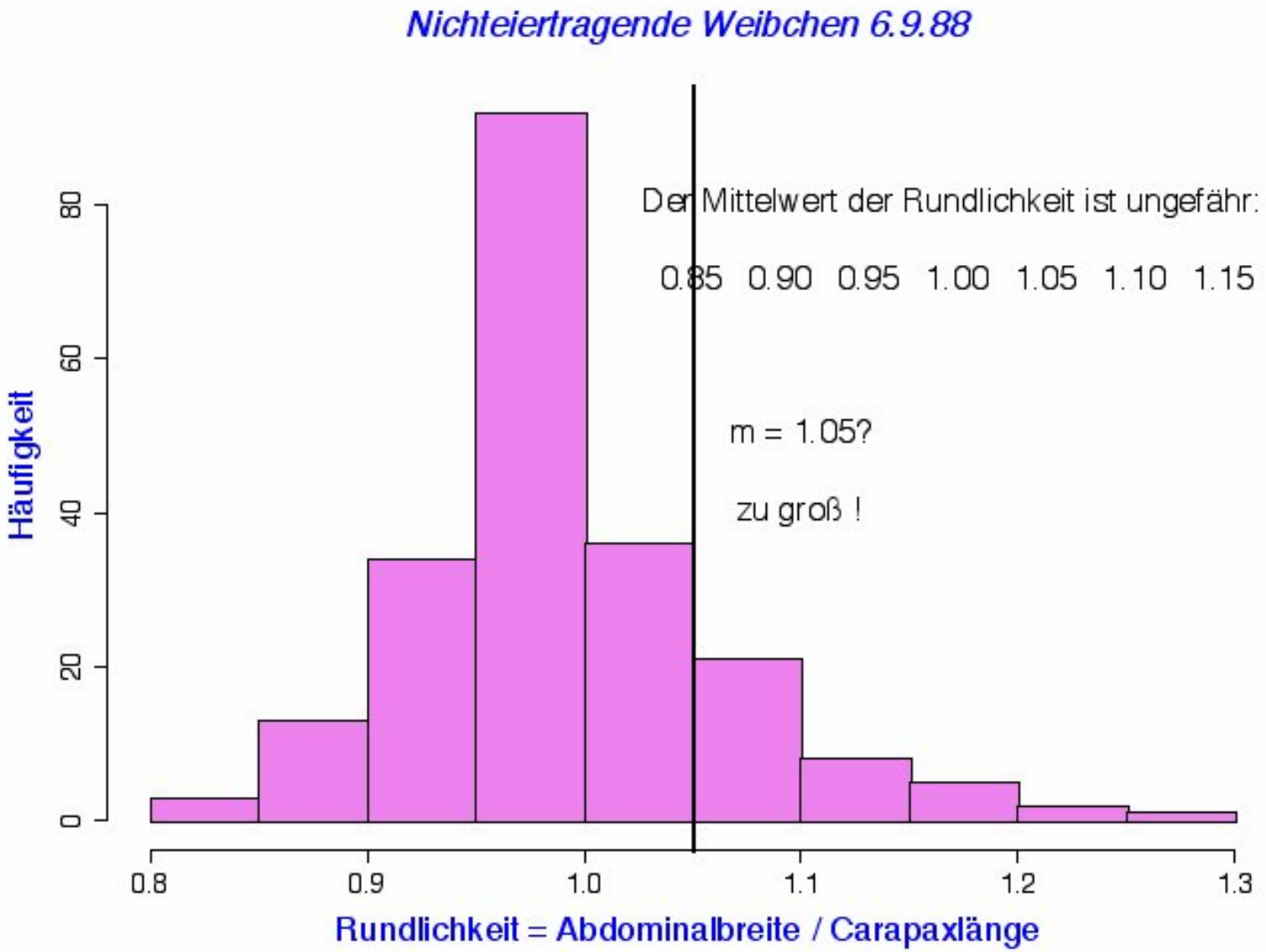

Nichteiertragende Weibchen 6.9.88
Häufigkeit
0
20
40
60
80
0.8
0.9
1.0
1.1
1.2
1.3
Rundlichkeit = Abdominalbreite / Carapaxlänge

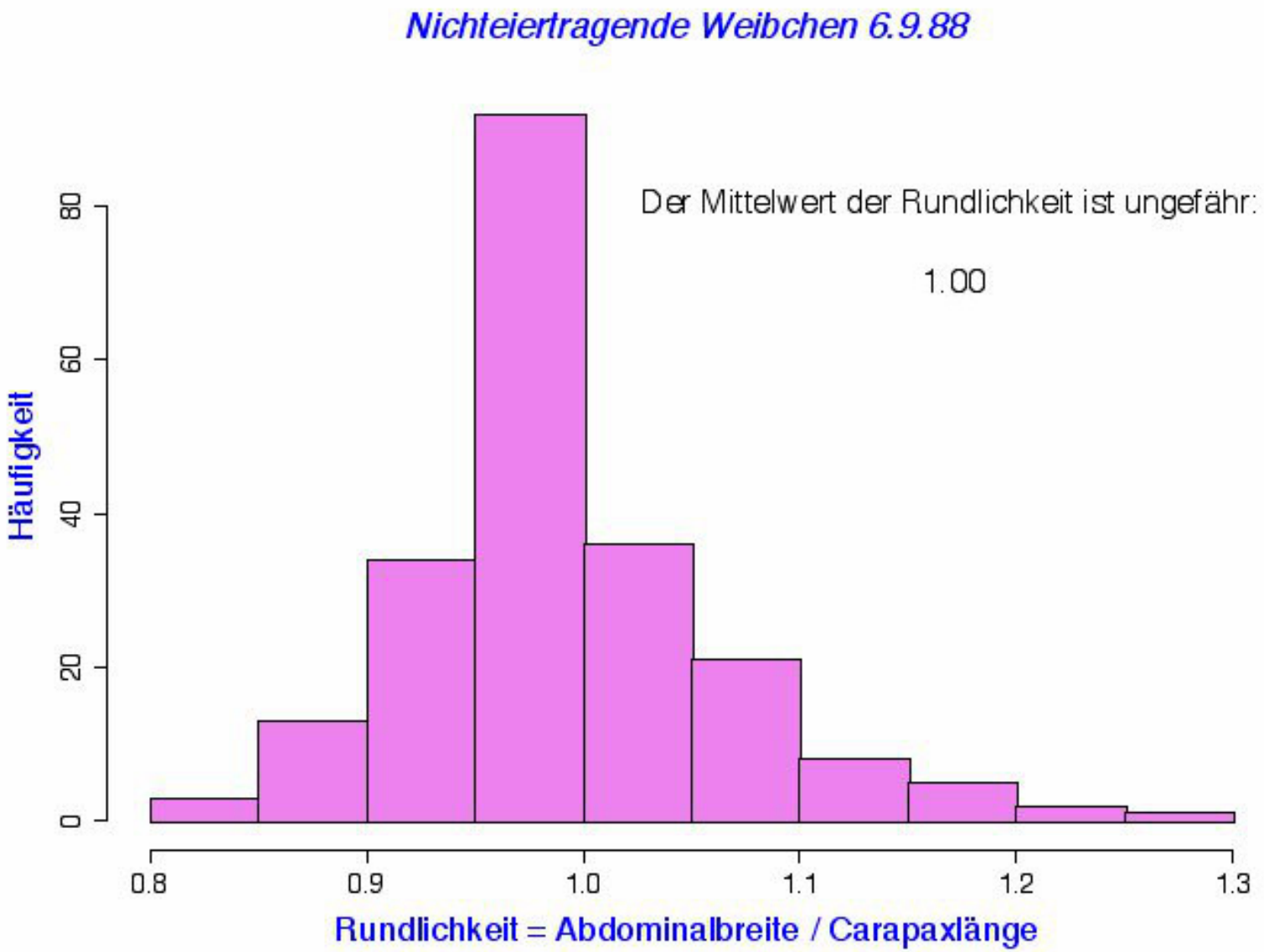

Nichteiertragende Weibchen 6.9.88
Der Mittelwert der Rundlichkeit ist ungefähr:
0.85 0.90 0.95 1.00 1.05 1.10 1.15
Häufigkeit
0
20
40
60
80
0.8
0.9
1.0
1.1
1.2
1.3
Rundlichkeit = Abdominalbreite / Carapaxlänge

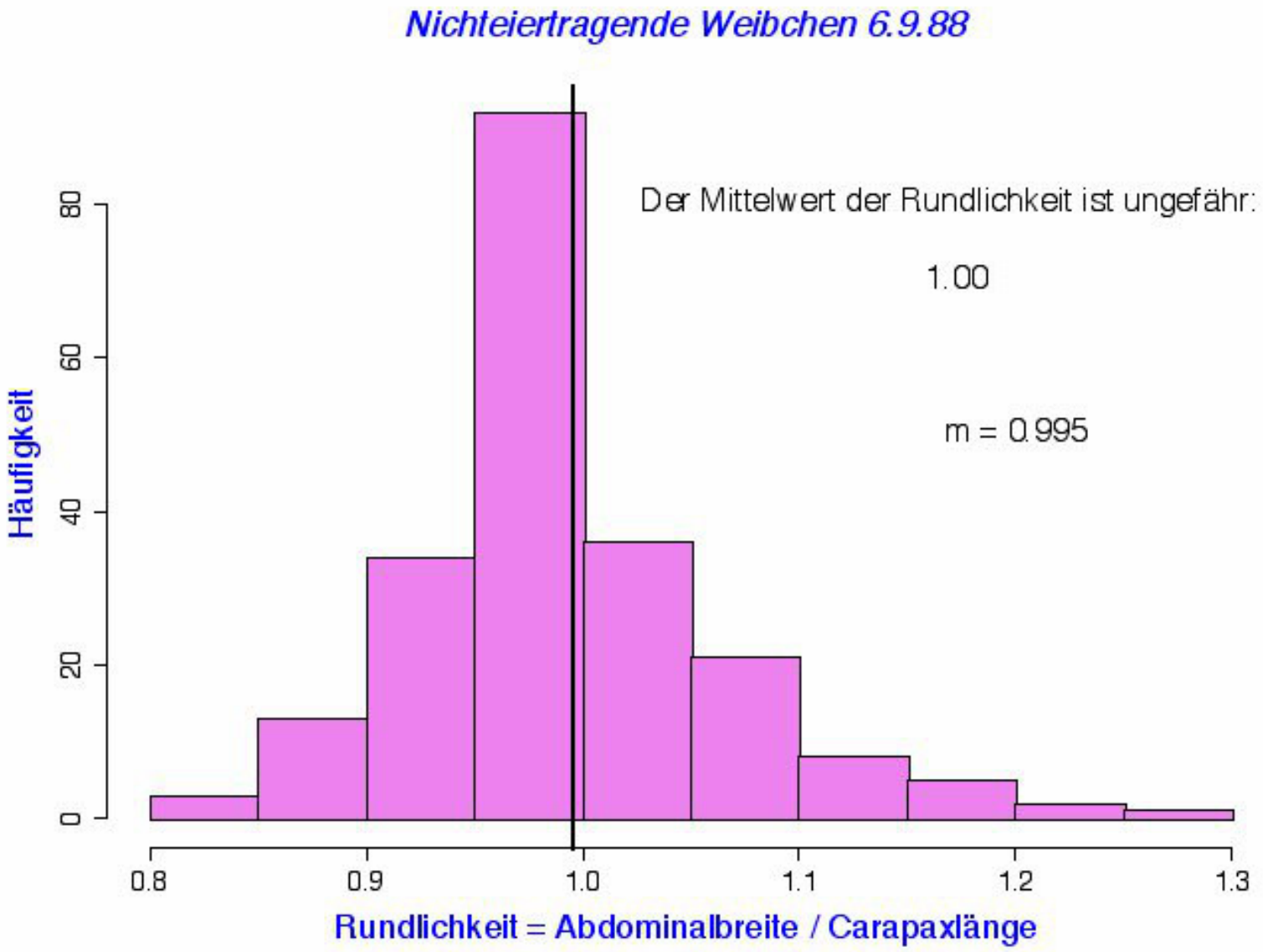

Nichteiertragende Weibchen 6.9.88
m = 0.95?
Der Mittelwert der Rundlichkeit ist ungefähr:
0.85 0.90 0.95 1.00 1.05 1.10 1.15
Häufigkeit
0
20
40
60
80
0.8
0.9
1.0
1.1
1.2
1.3
Rundlichkeit = Abdominalbreite / Carapaxlänge


Nichteiertragende Weibchen 6.9.88
m = 0.95?
zu klein !
Der Mittelwert der Rundlichkeit ist ungefähr:
0.85 0.90 0.95 1.00 1.05 1.10 1.15
Häufigkeit
0
20
40
60
80
0.8
0.9
1.0
1.1
1.2
1.3
Rundlichkeit = Abdominalbreite / Carapaxlänge


Nichteiertragende Weibchen 6.9.88
Der Mittelwert der Rundlichkeit ist ungefähr:
0.85 0.90 0.95 1.00 1.05 1.10 1.15
m = 1.05?
Häufigkeit
0
20
40
60
80
0.8
0.9
1.0
1.1
1.2
1.3
Rundlichkeit = Abdominalbreite / Carapaxlänge

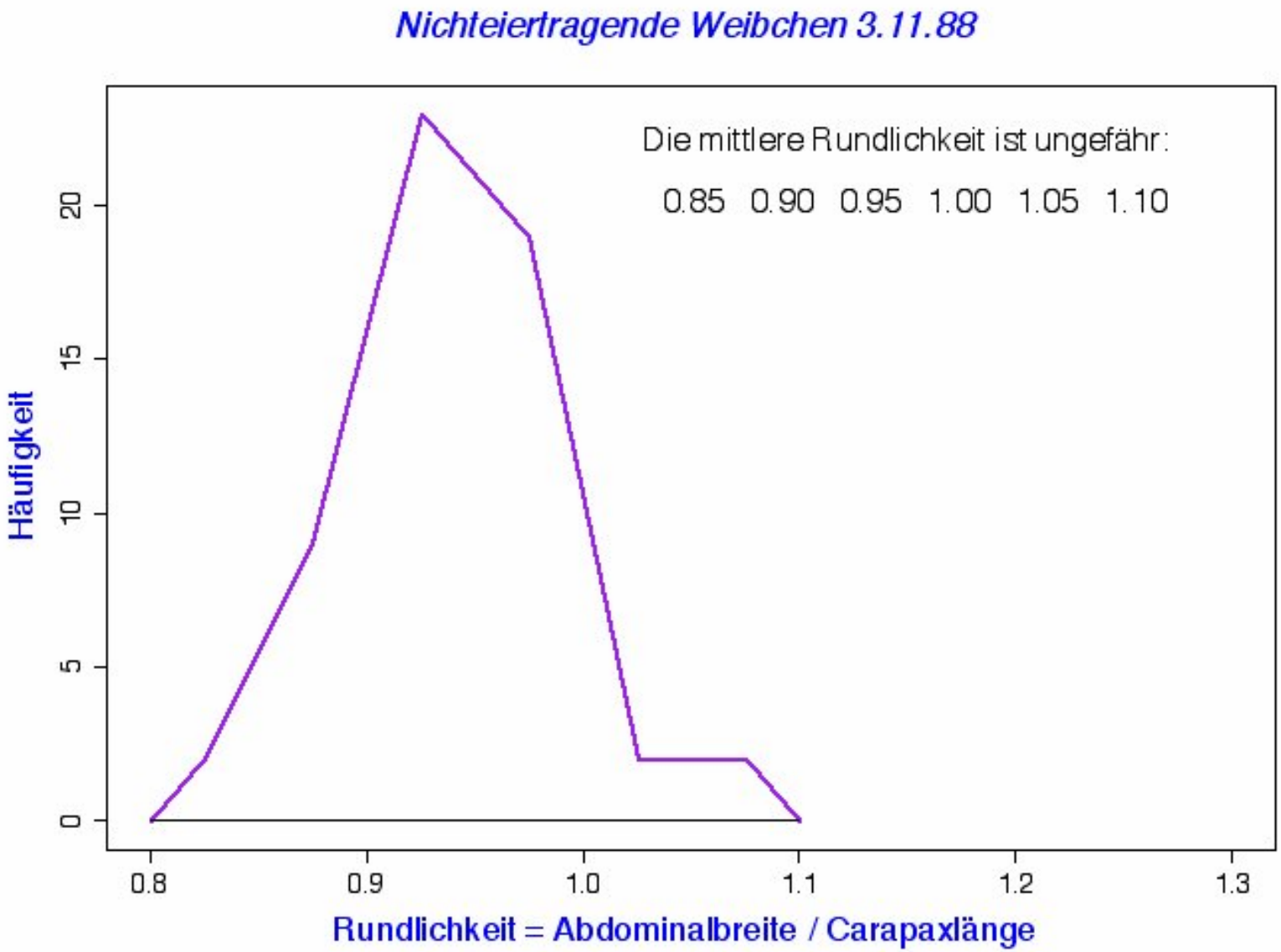

Nichteiertragende Weibchen 6.9.88
Der Mittelwert der Rundlichkeit ist ungefähr:
0.85 0.90 0.95 1.00 1.05 1.10 1.15
m = 1.05?
zu groß !
Häufigkeit
0
20
40
60
80
0.8
0.9
1.0
1.1
1.2
1.3
Rundlichkeit = Abdominalbreite / Carapaxlänge

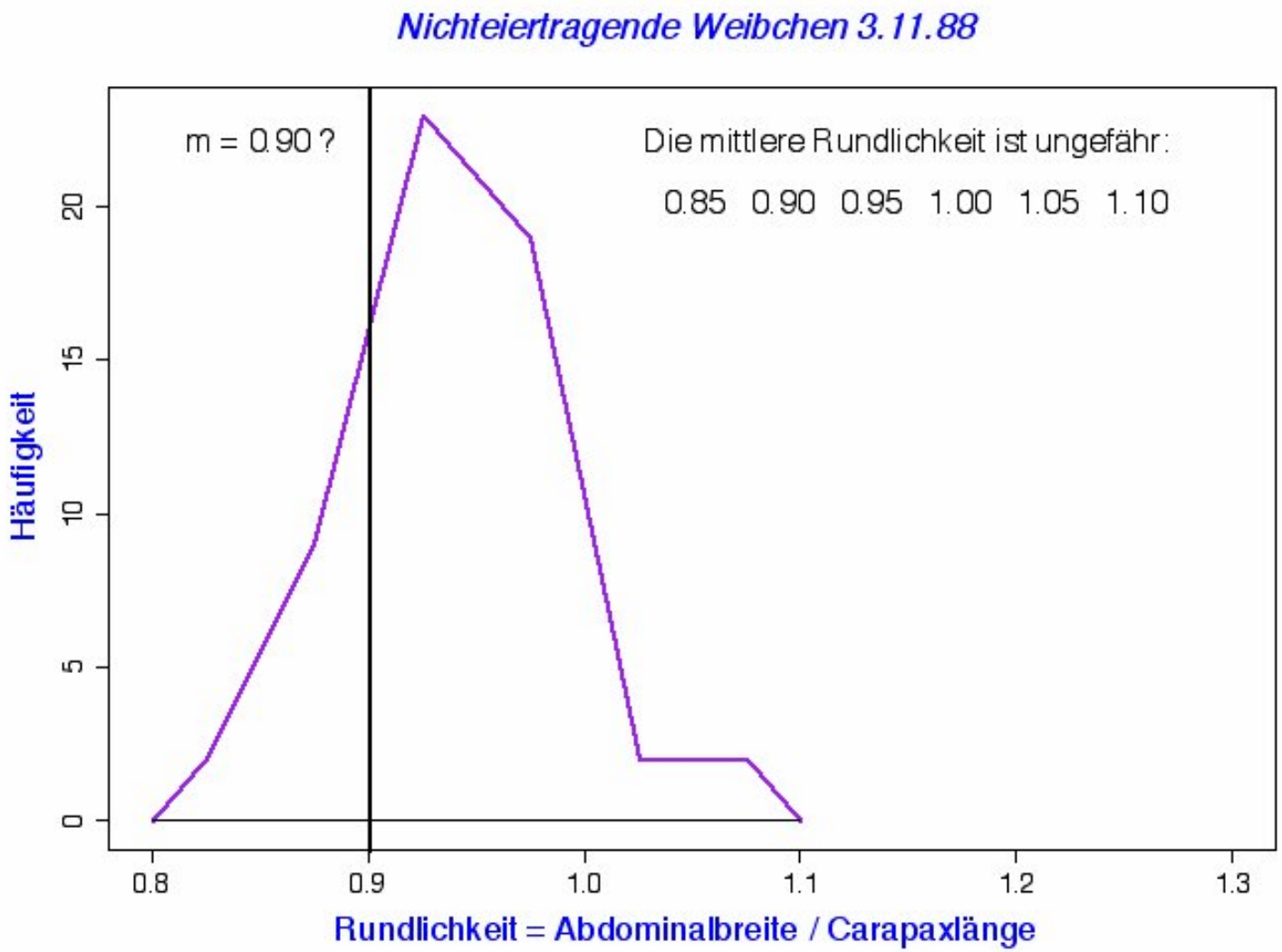

Nichteiertragende Weibchen 6.9.88
Der Mittelwert der Rundlichkeit ist ungefähr:
1.00
Häufigkeit
0
20
40
60
80
0.8
0.9
1.0
1.1
1.2
1.3
Rundlichkeit = Abdominalbreite / Carapaxlänge

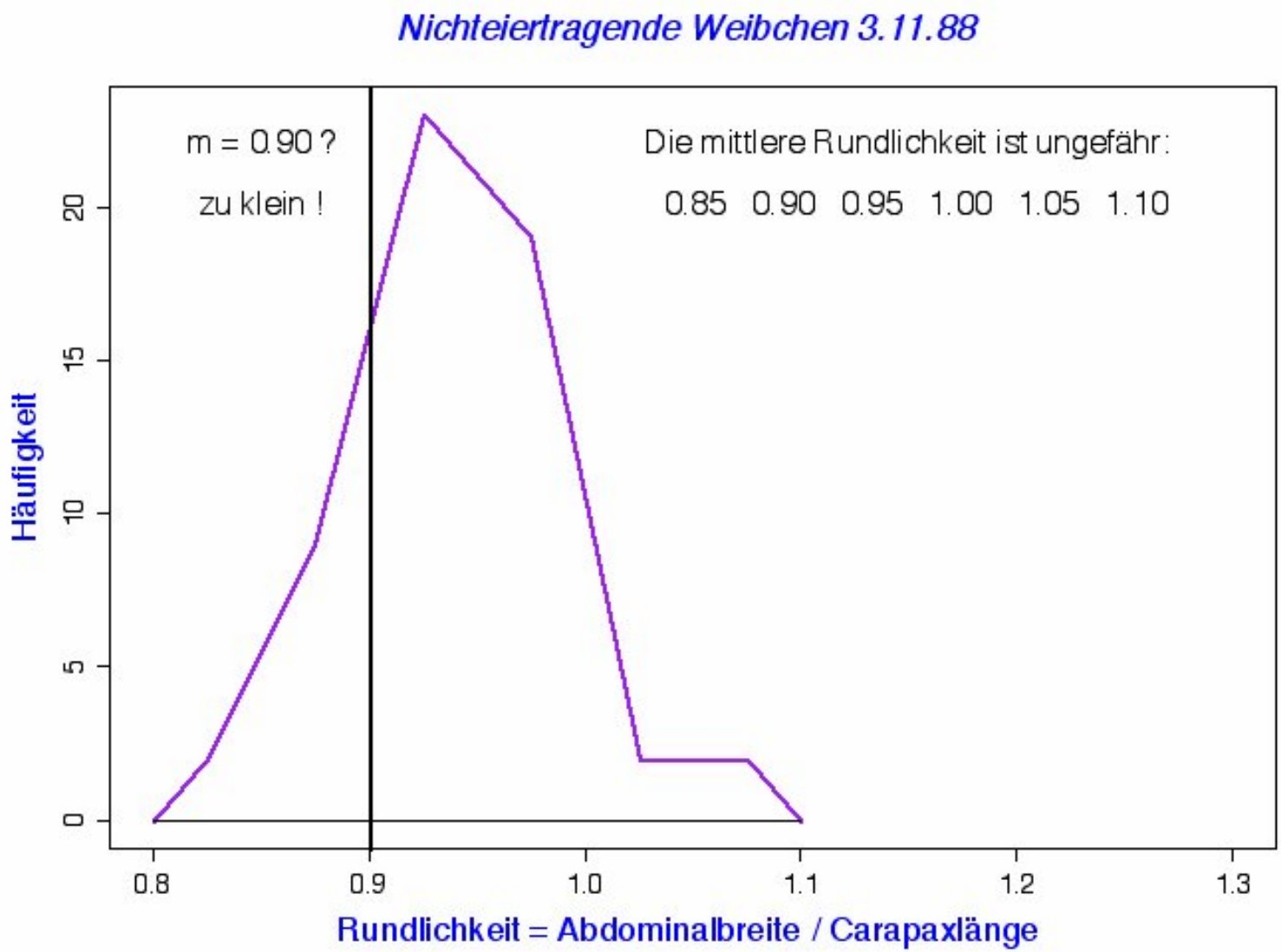

Nichteiertragende Weibchen 6.9.88
Der Mittelwert der Rundlichkeit ist ungefähr:
1.00
m = 0.995
Häufigkeit
0
20
40
60
80
0.8
0.9
1.0
1.1
1.2
1.3
Rundlichkeit = Abdominalbreite / Carapaxlänge

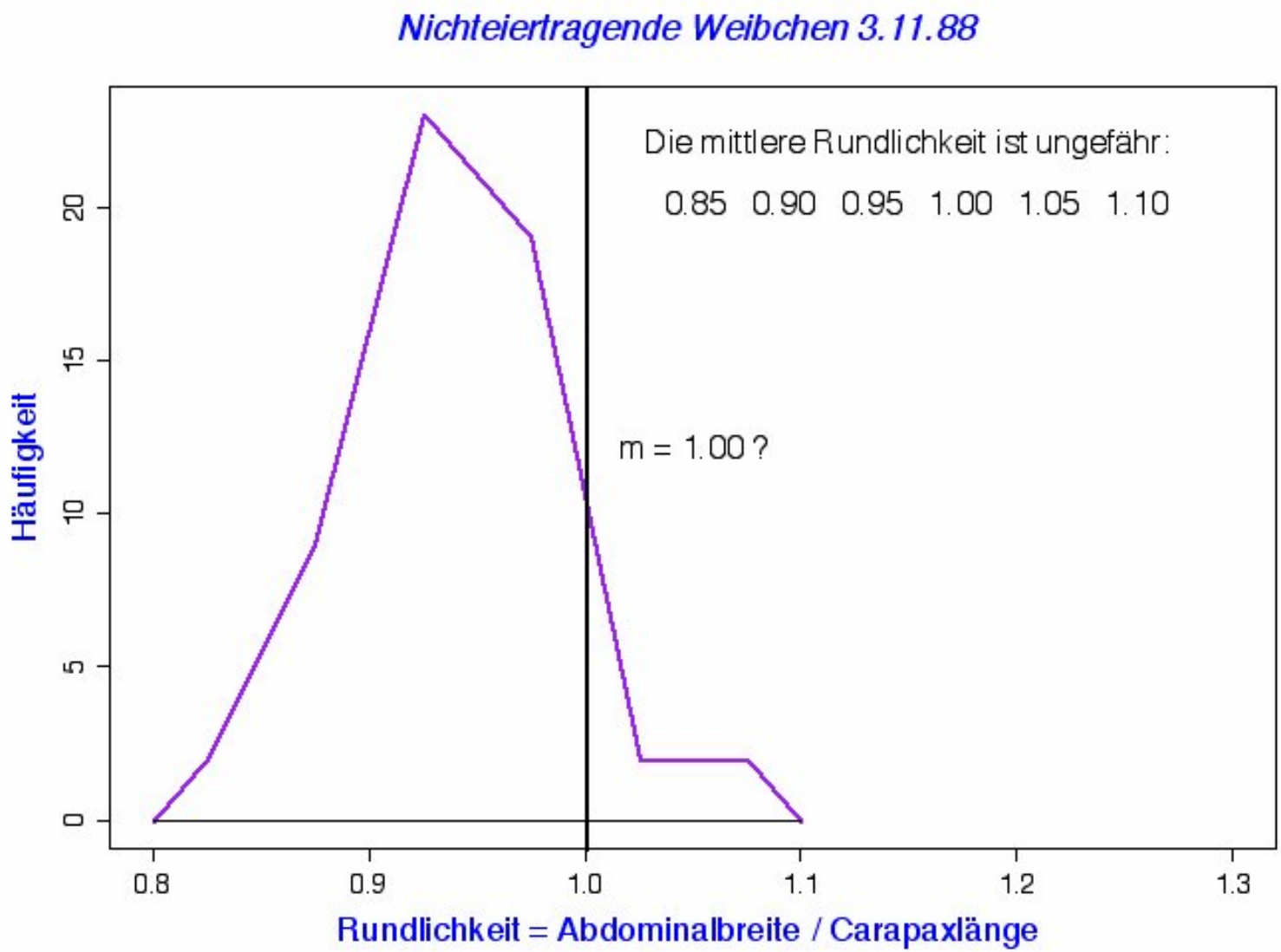
Beispiel:

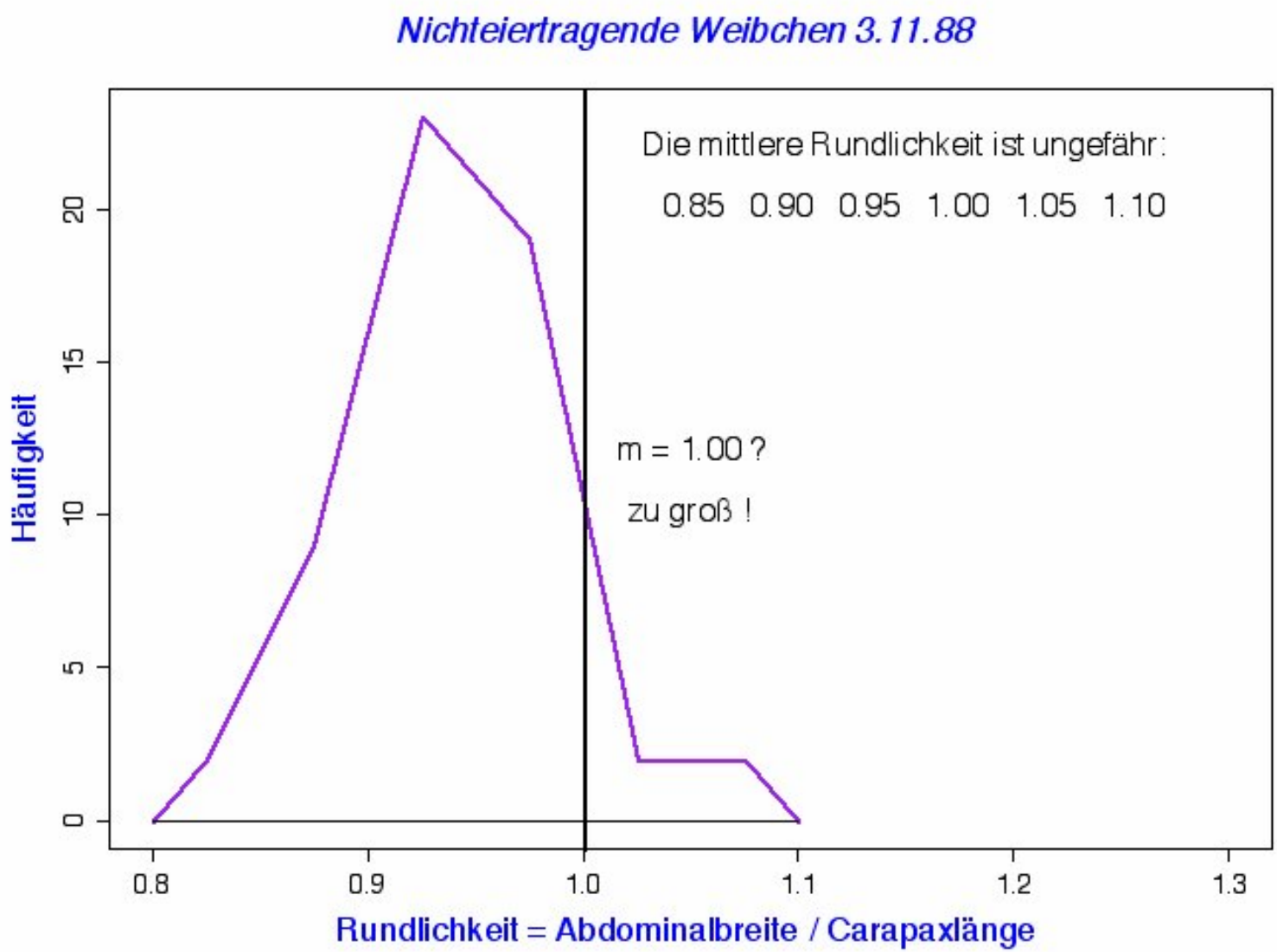
3.11.88

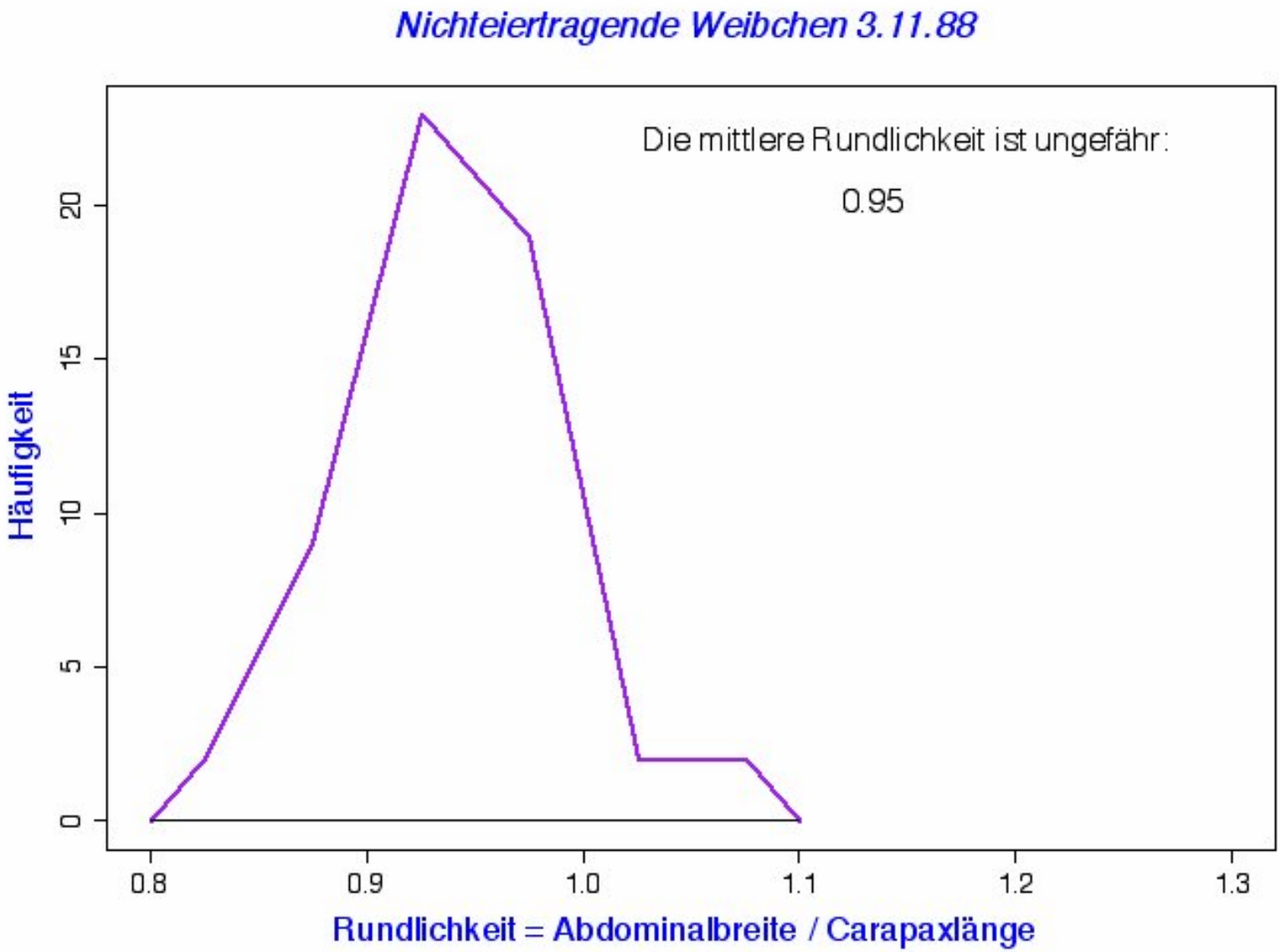

Nichteiertragende Weibchen 3.11.88
Die mittlere Rundlichkeit ist ungefähr:
0.85 0.90 0.95 1.00 1.05 1.10
Häufigkeit
0
5
10
15
20
0.8
0.9
1.0
1.1
1.2
1.3
Rundlichkeit = Abdominalbreite / Carapaxlänge

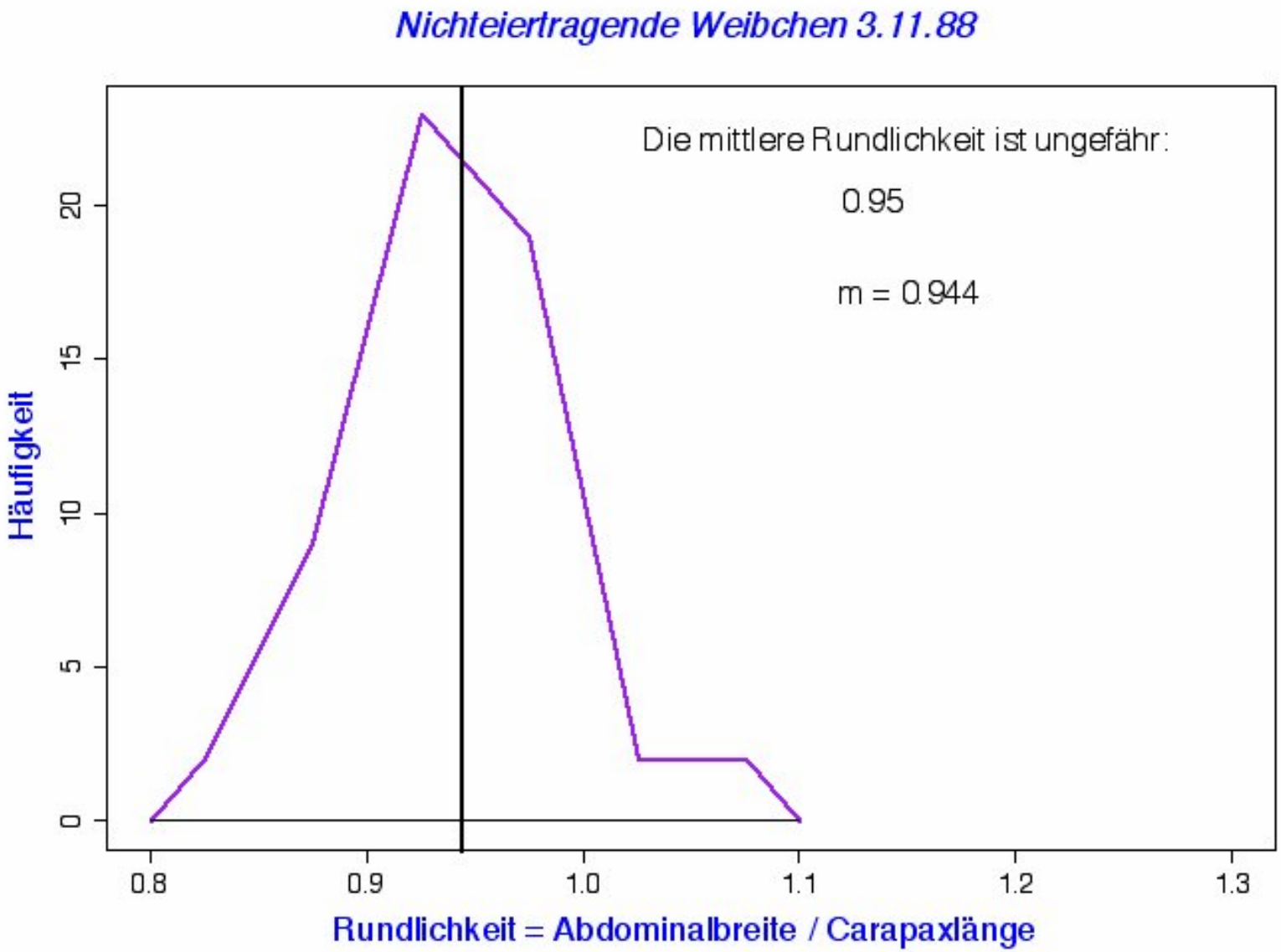

Nichteiertragende Weibchen 3.11.88
m = 0.90 ?
Die mittlere Rundlichkeit ist ungefähr:
0.85 0.90 0.95 1.00 1.05 1.10
Häufigkeit
0
5
10
15
20
0.8
0.9
1.0
1.1
1.2
1.3
Rundlichkeit = Abdominalbreite / Carapaxlänge


Nichteiertragende Weibchen 3.11.88
m = 0.90 ?
zu klein !
Die mittlere Rundlichkeit ist ungefähr:
0.85 0.90 0.95 1.00 1.05 1.10
Häufigkeit
0
5
10
15
20
0.8
0.9
1.0
1.1
1.2
1.3
Rundlichkeit = Abdominalbreite / Carapaxlänge


Nichteiertragende Weibchen 3.11.88
Die mittlere Rundlichkeit ist ungefähr:
0.85 0.90 0.95 1.00 1.05 1.10
m = 1.00 ?
Häufigkeit
0
5
10
15
20
0.8
0.9
1.0
1.1
1.2
1.3
Rundlichkeit = Abdominalbreite / Carapaxlänge


Nichteiertragende Weibchen 3.11.88
Die mittlere Rundlichkeit ist ungefähr:
0.85 0.90 0.95 1.00 1.05 1.10
m = 1.00 ?
zu groß !
Häufigkeit
0
5
10
15
20
0.8
0.9
1.0
1.1
1.2
1.3
Rundlichkeit = Abdominalbreite / Carapaxlänge


Nichteiertragende Weibchen 3.11.88
Die mittlere Rundlichkeit ist ungefähr:
0.95
Häufigkeit
0
5
10
15
20
0.8
0.9
1.0
1.1
1.2
1.3
Rundlichkeit = Abdominalbreite / Carapaxlänge


Nichteiertragende Weibchen 3.11.88
Die mittlere Rundlichkeit ist ungefähr:
0.95
m = 0.944
Häufigkeit
0
5
10
15
20
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 $n-1$ berechnen?

Die Standardabweichung σ eines Zufallsexperiments mit n gleichwahrscheinlichen Ausgängen $x_1, \ldots, x_n$ (z.B. Würfelwurf) ist definiert durch

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

Mit n oder $n-1$ berechnen?

Die Standardabweichung σ eines Zufallsexperiments mit n gleichwahrscheinlichen Ausgängen $x_1, \ldots, x_n$ (z.B. Würfelwurf) ist definiert durch

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

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 σ

Wir werden sehen:

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

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

so hat die *Zufallsvariable*

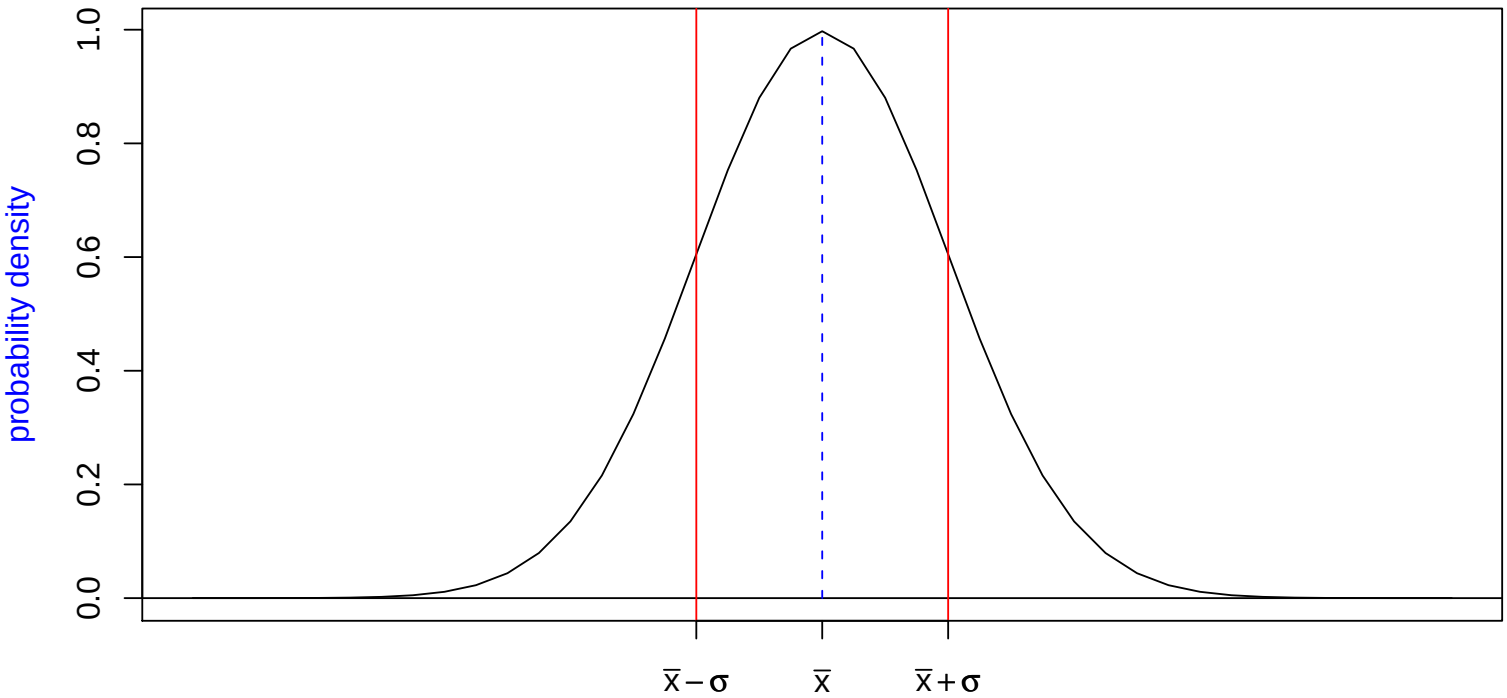
$$S^2 := \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \overline{X})^2$$

die Eigenschaft

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

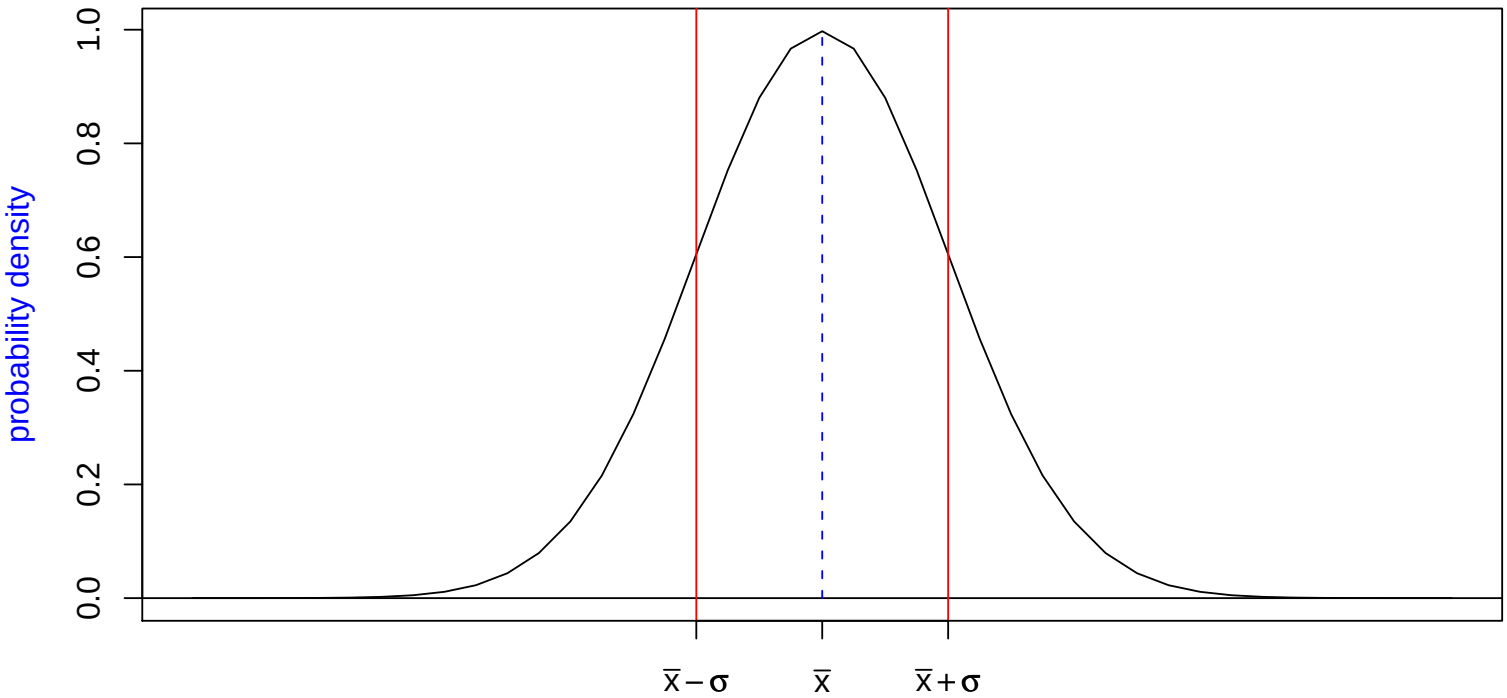
Faustregel für die Standardabweichung

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



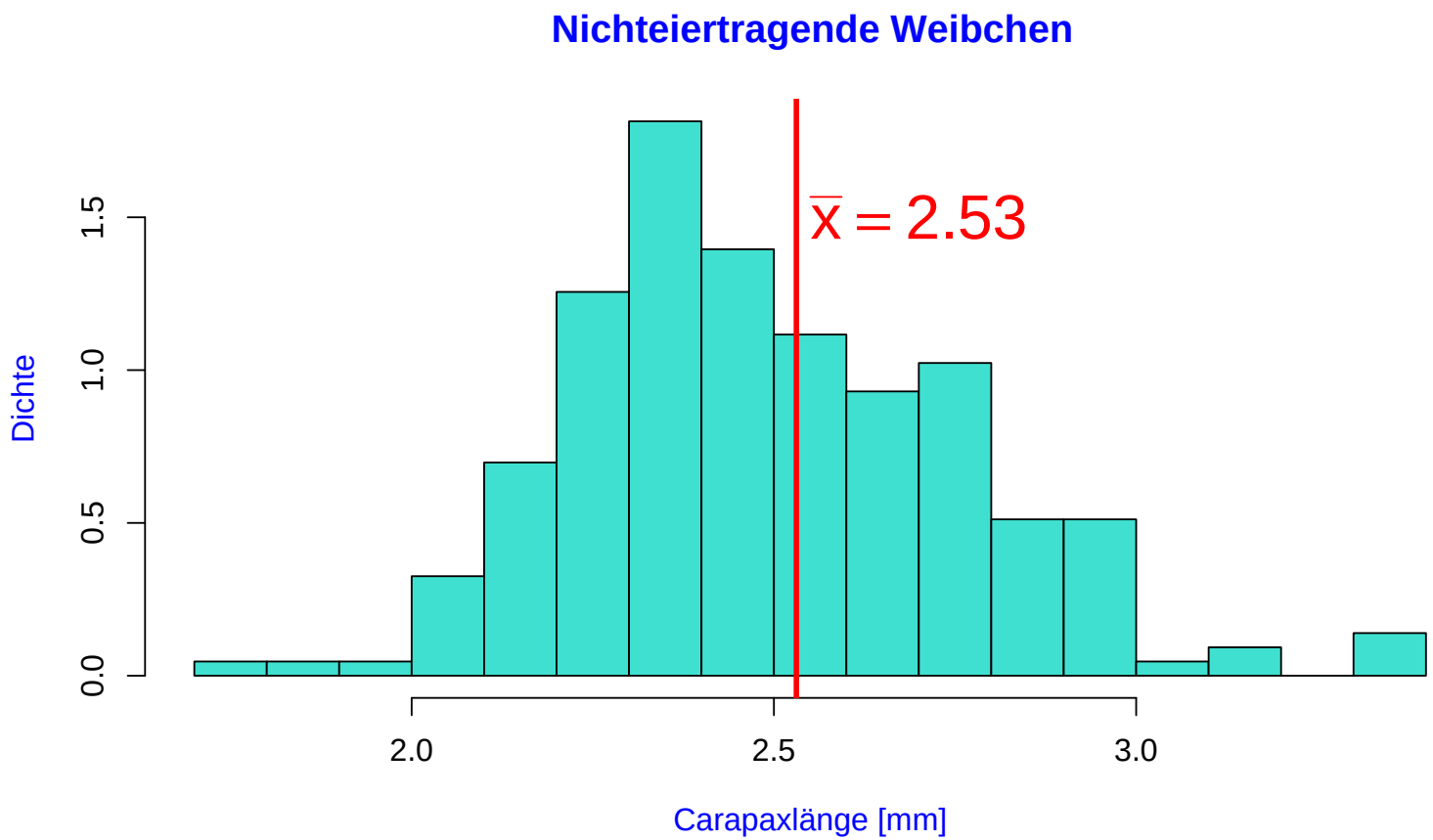
Faustregel für die Standardabweichung

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

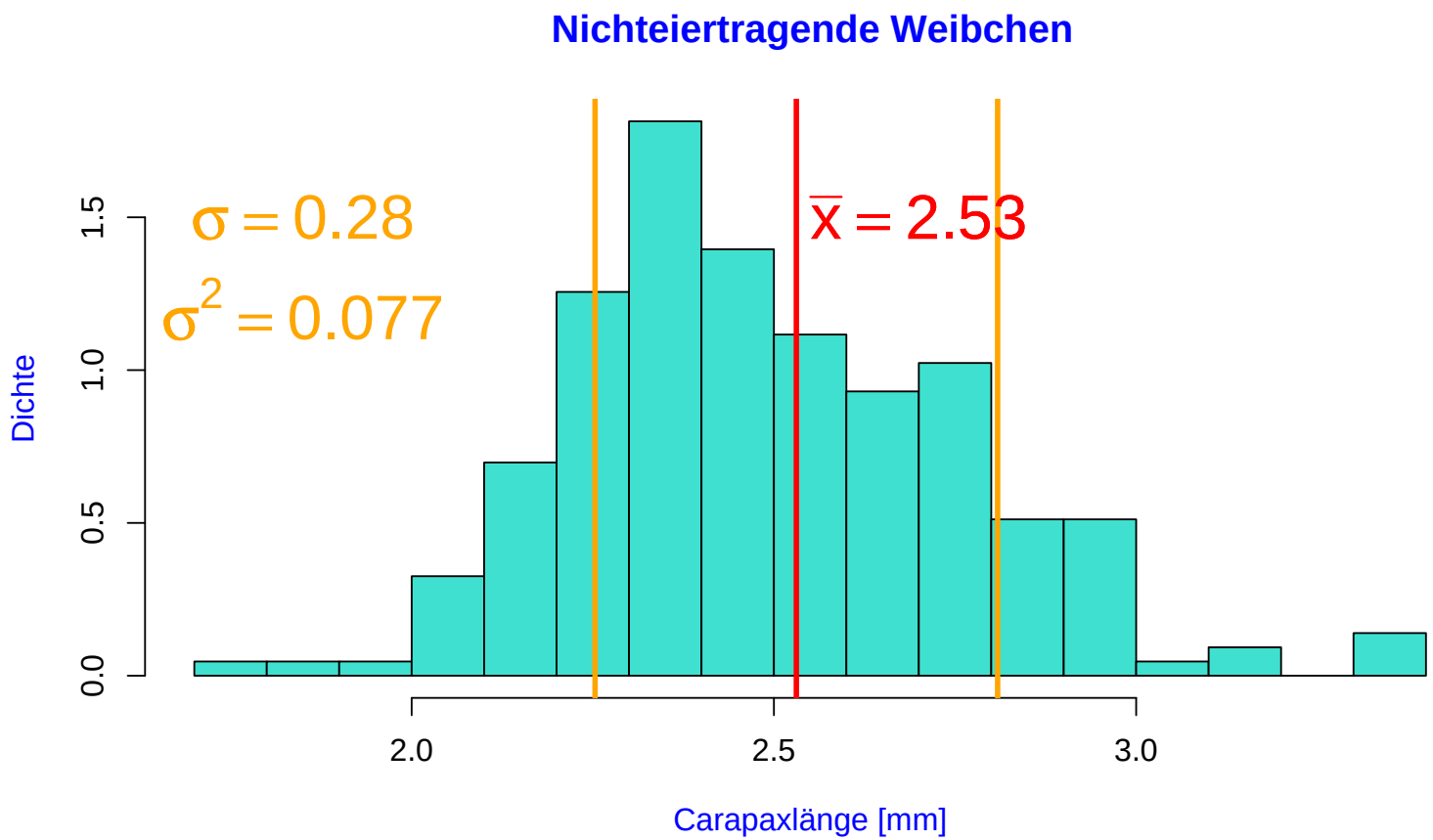


Oft kann man so die Standardabweichung „mit bloßem Auge“ abschätzen.

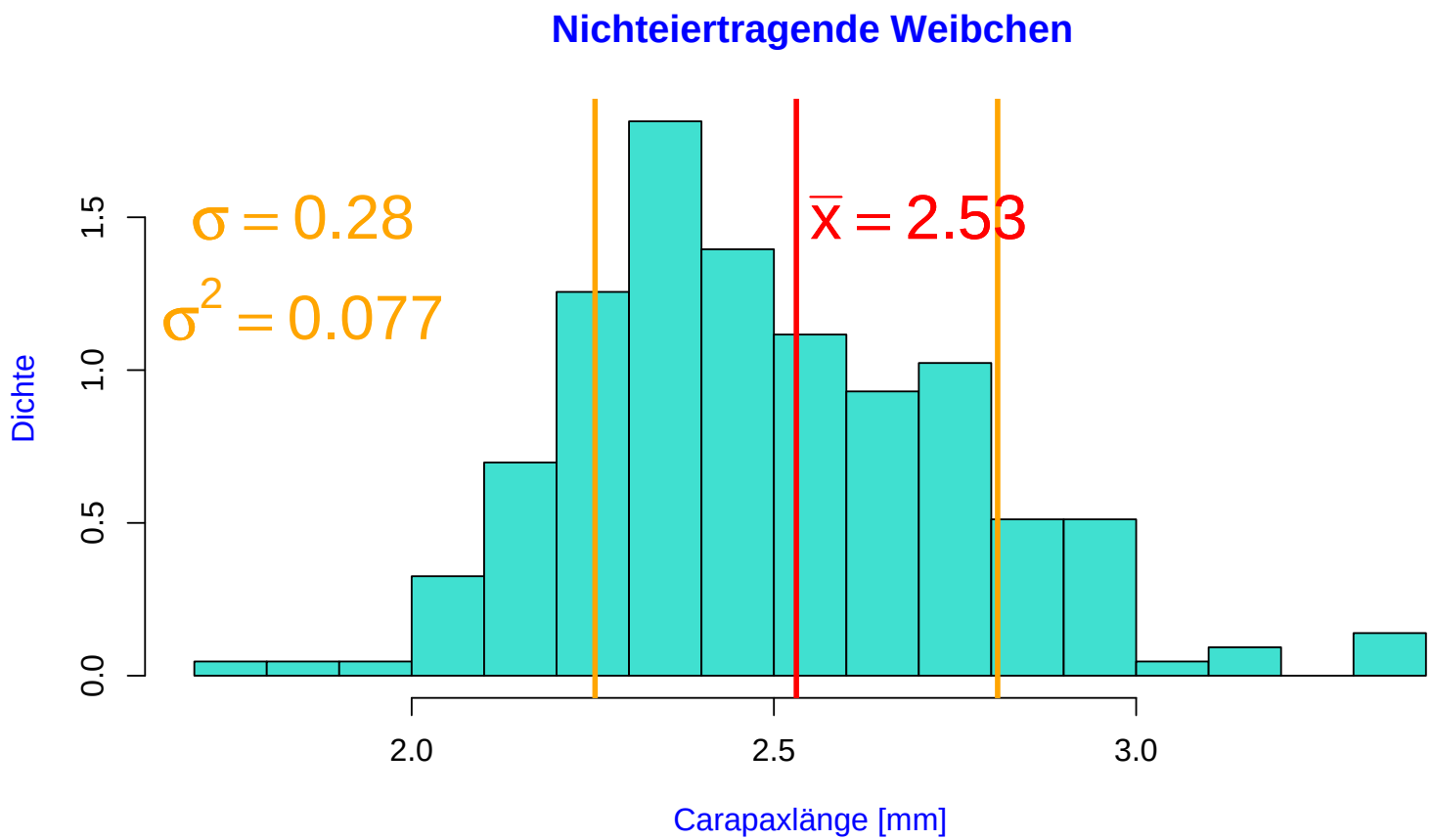
Standardabweichung der Carapaxlängen nichteiertragender Weibchen vom 6.9.88



Standardabweichung der Carapaxlängen nichteiertragender Weibchen vom 6.9.88



Standardabweichung der Carapaxlängen nichteiertragender Weibchen vom 6.9.88



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

Inhalt

Mittelwert und Standardabweichung. . .

- charakterisieren die Daten gut, falls deren Verteilung glockenförmig ist

Mittelwert und Standardabweichung. . .

- charakterisieren die Daten gut, falls deren Verteilung glockenförmig ist
- und müssen andernfalls mit Vorsicht interpretiert werden.

Mittelwert und Standardabweichung. . .

- charakterisieren die Daten gut, falls deren Verteilung glockenförmig ist
- und müssen andernfalls mit Vorsicht interpretiert werden.

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

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

Mittelwert und Standardabweichung. . .

- charakterisieren die Daten gut, falls deren Verteilung glockenförmig ist
- und müssen andernfalls mit Vorsicht interpretiert werden.

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

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

Im Folgenden verwenden wir zum Teil simulierte Daten, wenn die Originaldaten nicht verfügbar waren.
(Nehmen Sie also nicht alle Datenpunkte wörtlich.)

Inhalt

Bachstelzen fressen Dungfliegen

Räuber

Bachstelze (White Wagtail)
Motacilla alba alba

image (c) by Artur Mikołajewski

Beute

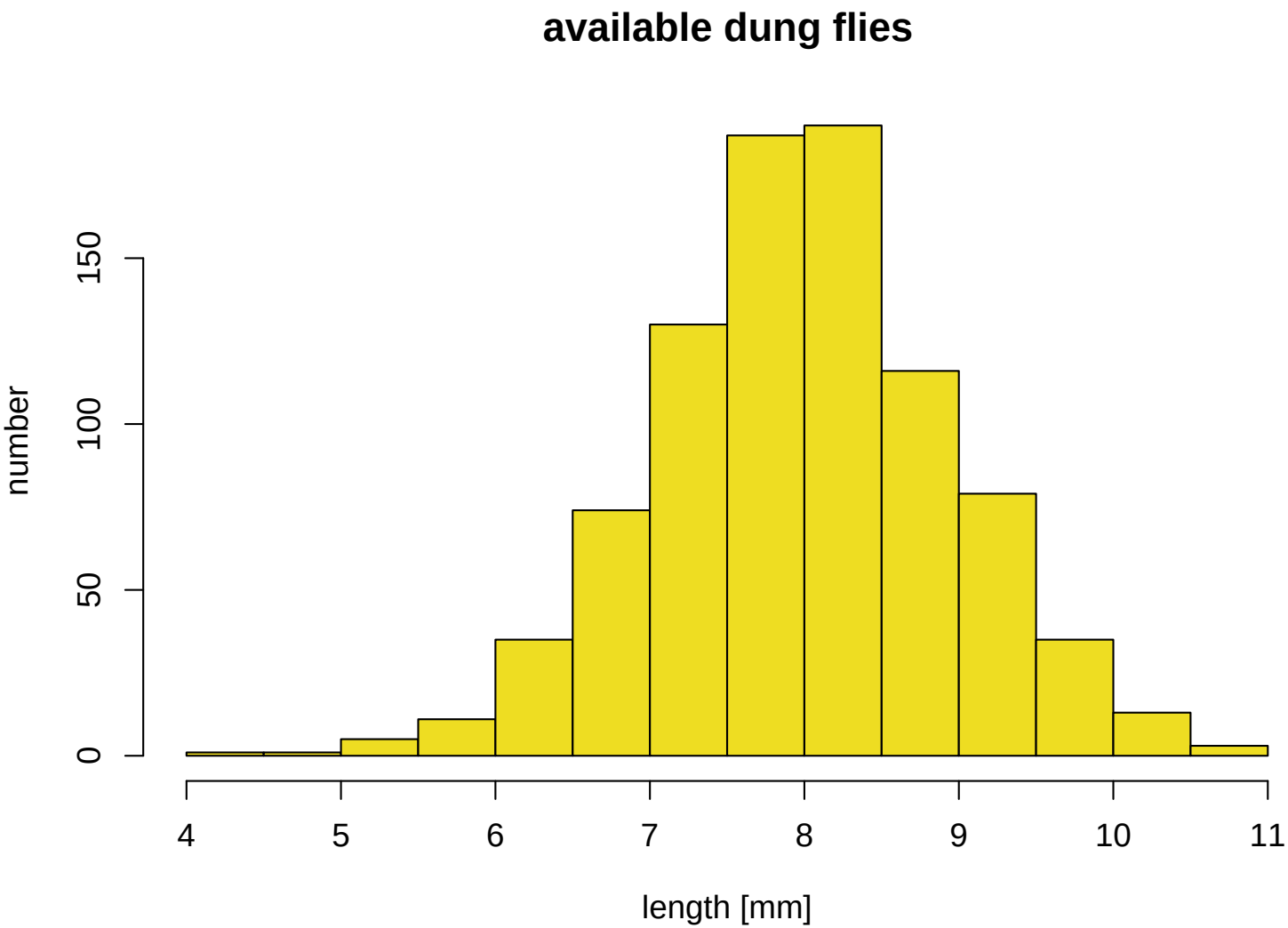
Gelbe Dungfliege
Scatophaga stercoraria

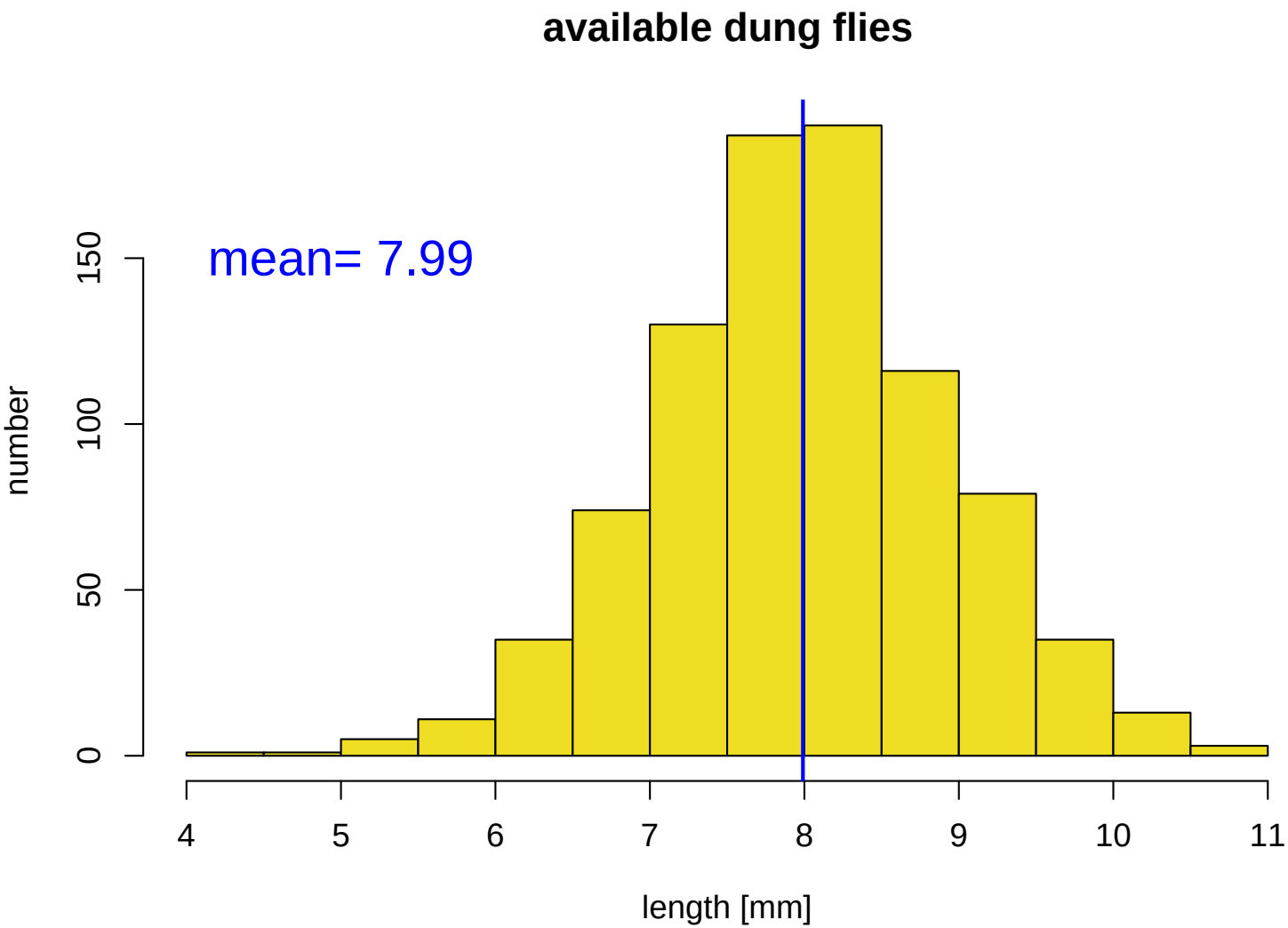
image (c) by Viatour Luc

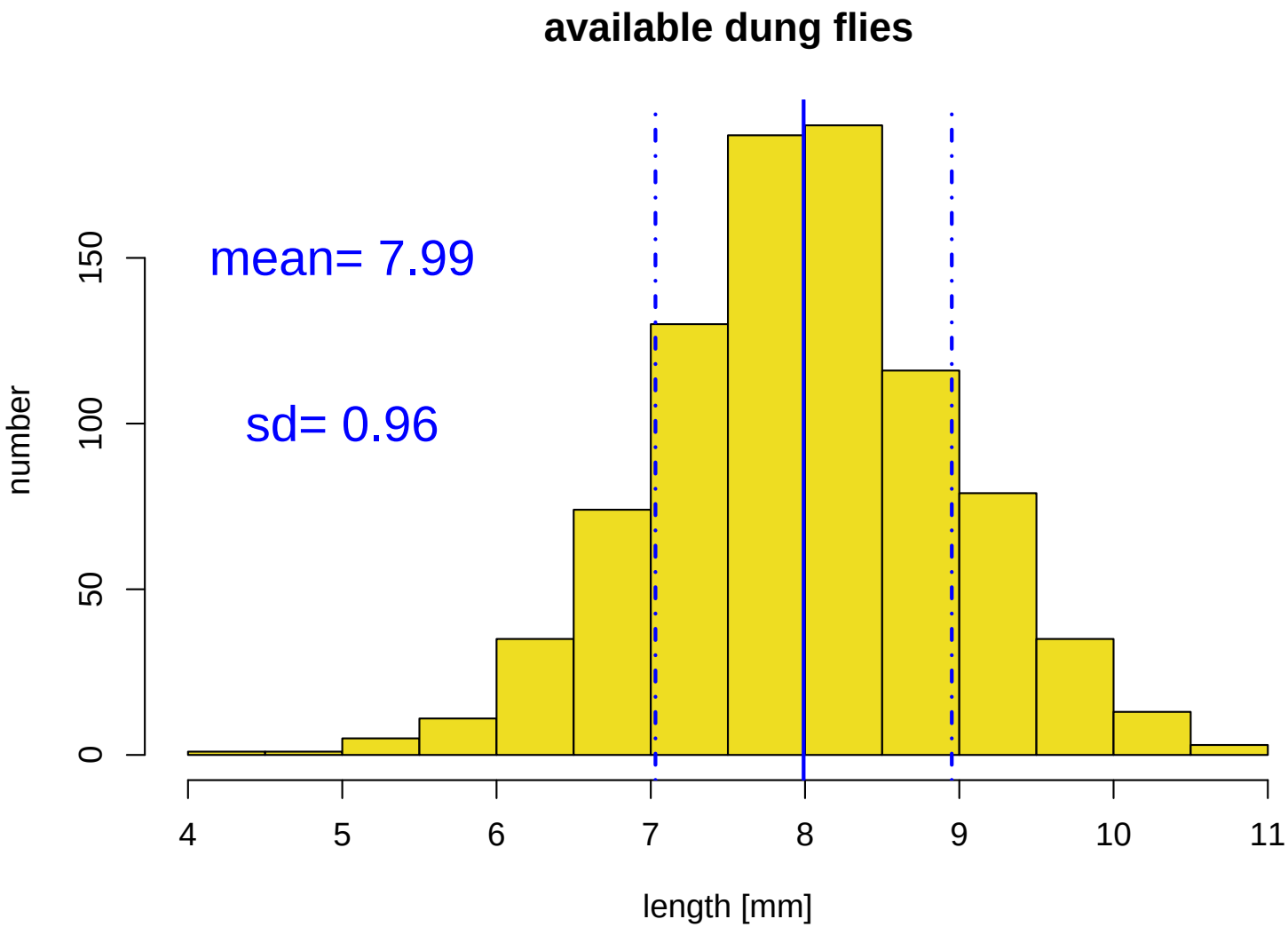
Vermutung

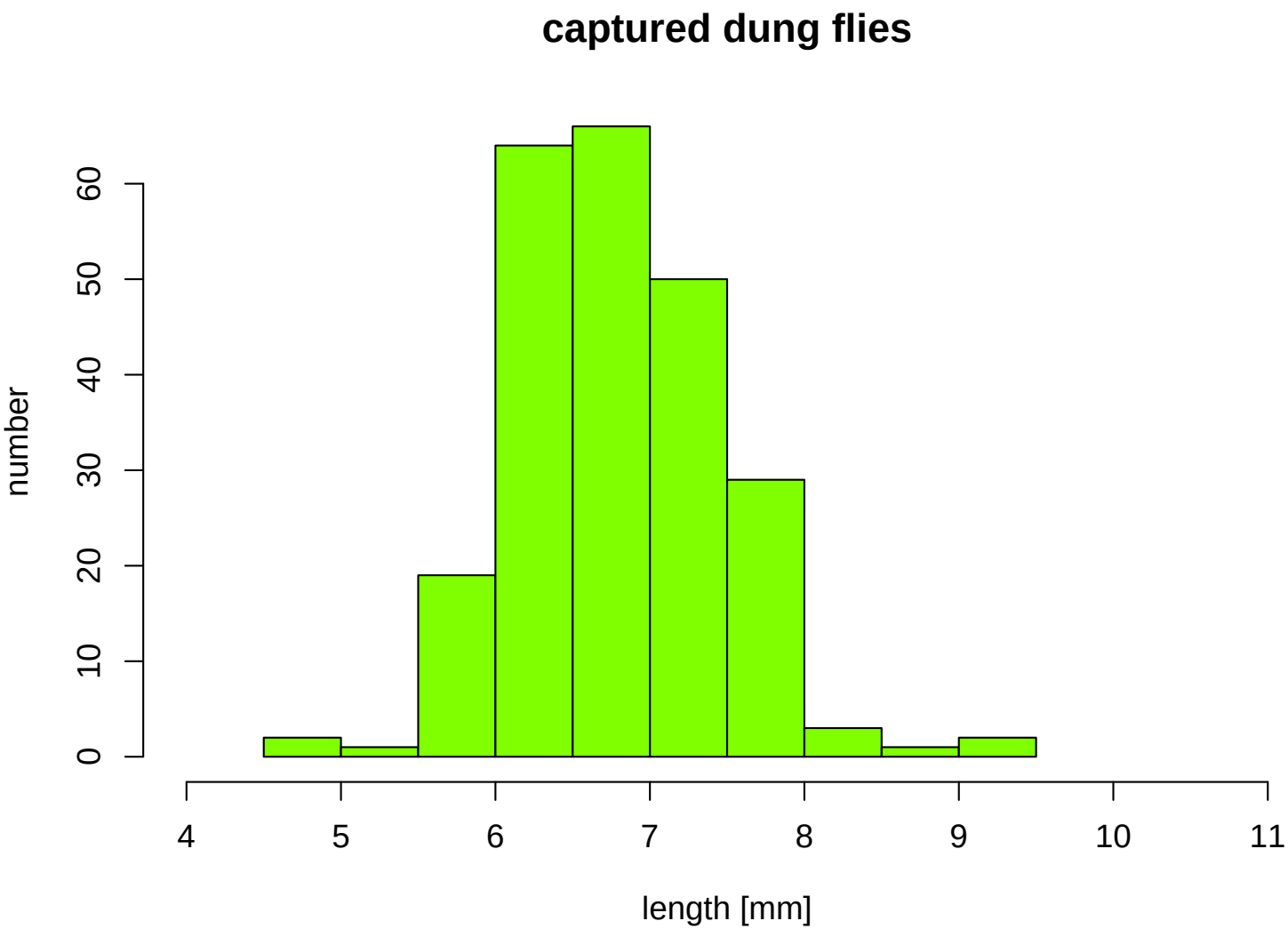
- Die Fliegen sind unterschiedlich groß
- Effizienz für die Bachstelze = Energiegewinn / Zeit zum Fangen und fressen
- Laborexperimente lassen vermuten, dass die Effizienz bei 7mm großen Fliegen maximal ist.

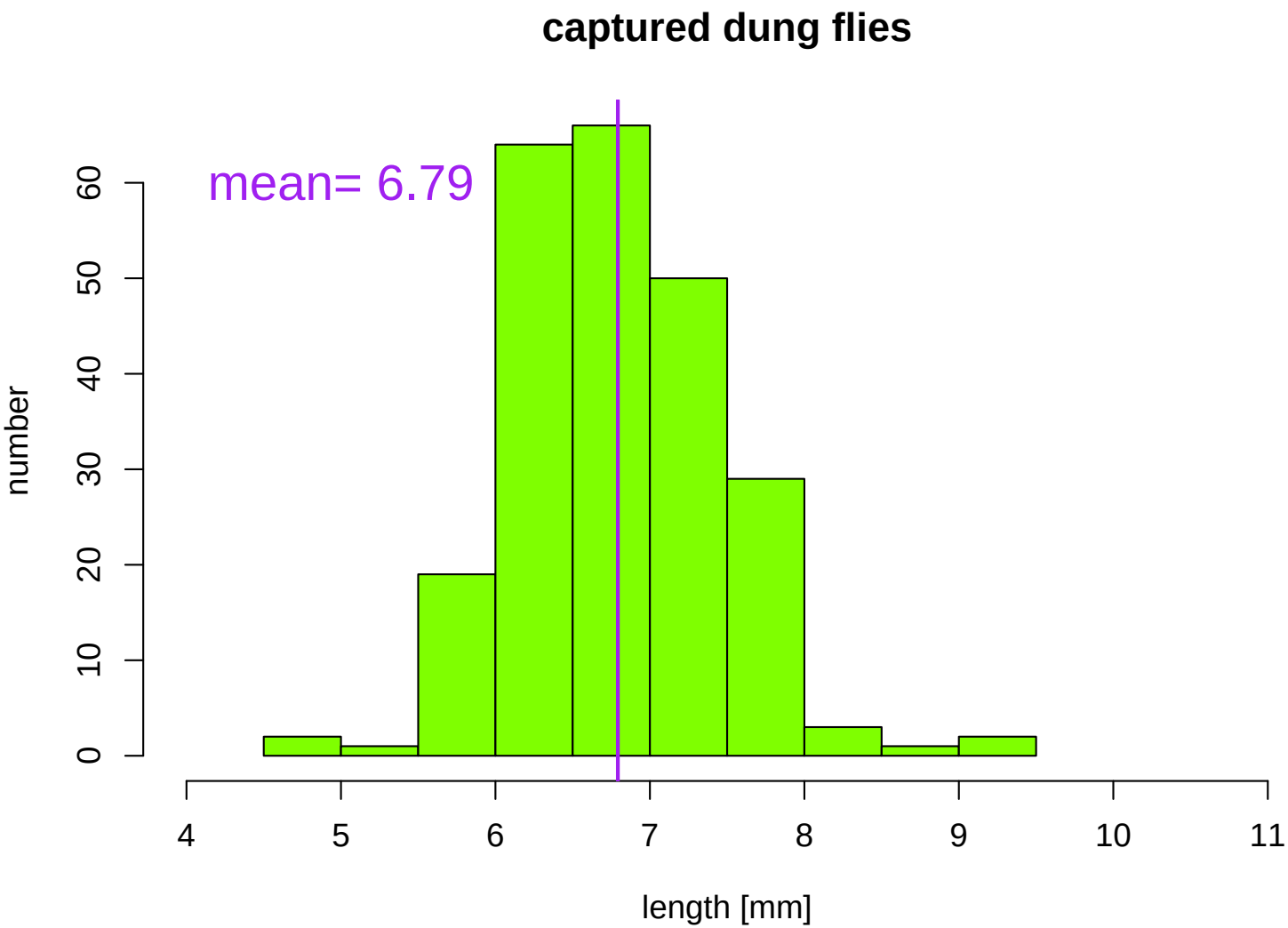
N.B. Davies.
Prey selection and social behaviour in wagtails (Aves: Motacillidae).
J. Anim. Ecol., 46:37–57, 1977.

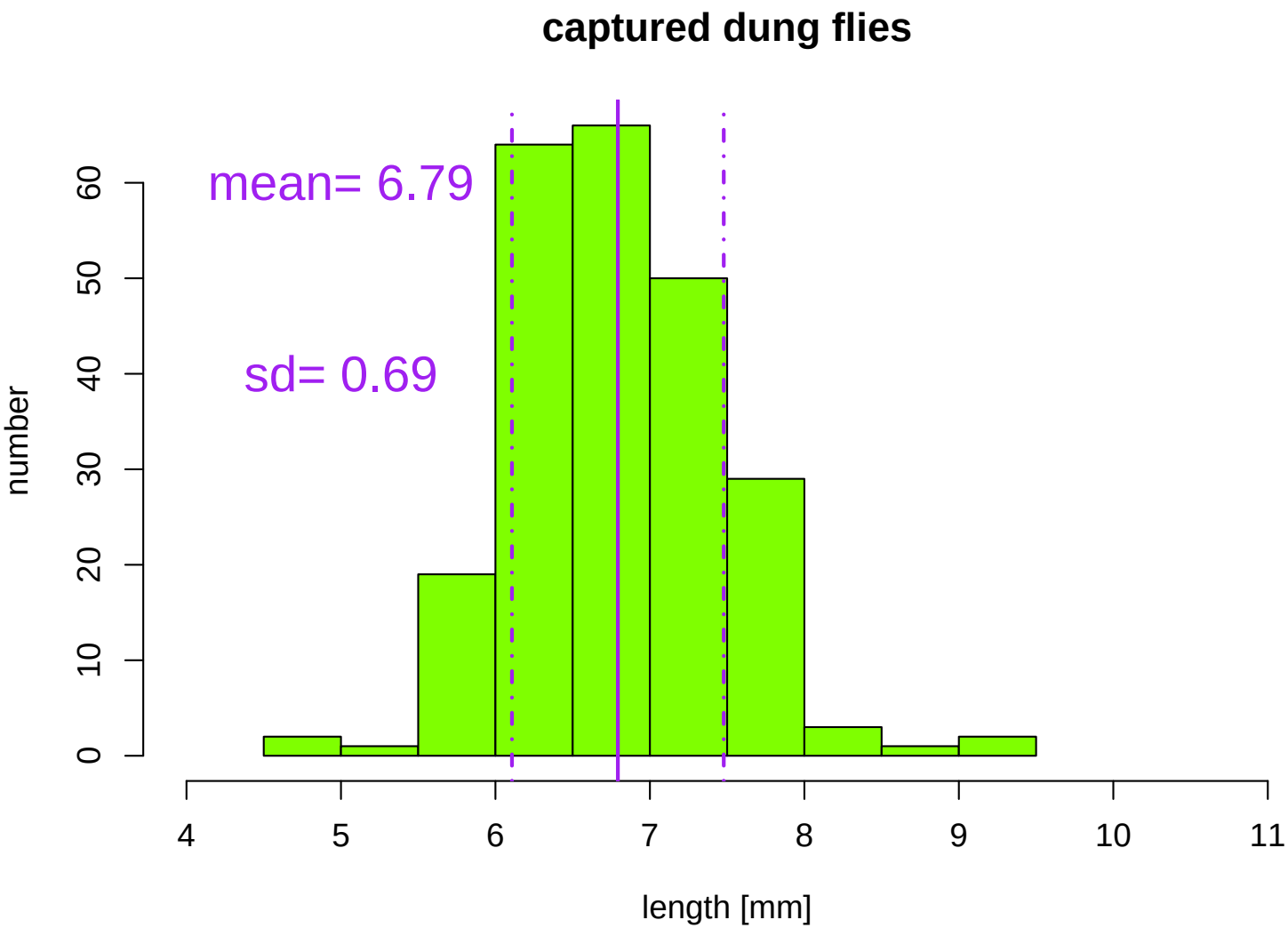

available dung flies
number
0
50
100
150
4
5
6
7
8
9
10
11
length [mm]

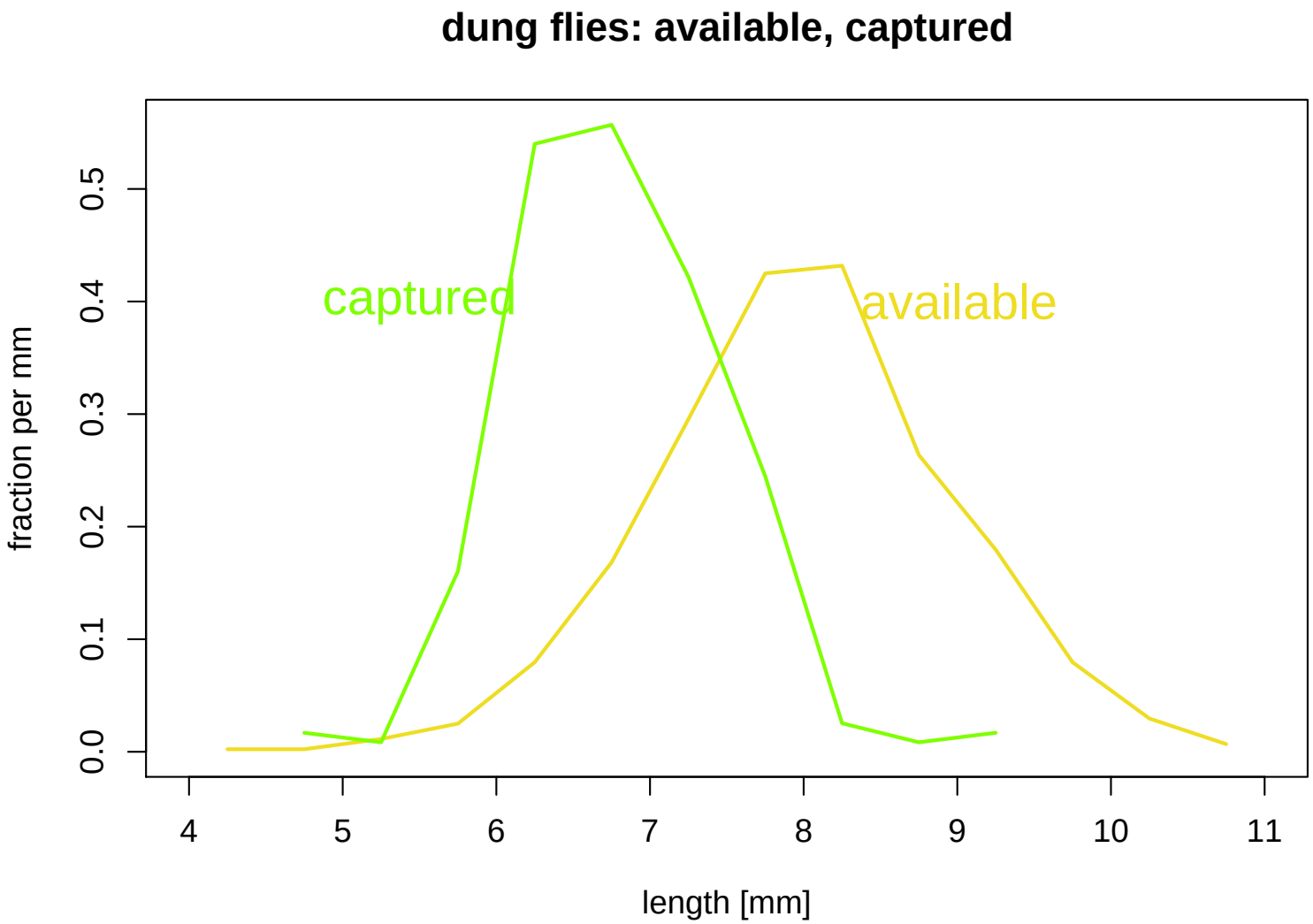

available dung flies
mean= 7.99
number
0
50
100
150
4
5
6
7
8
9
10
11
length [mm]


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]

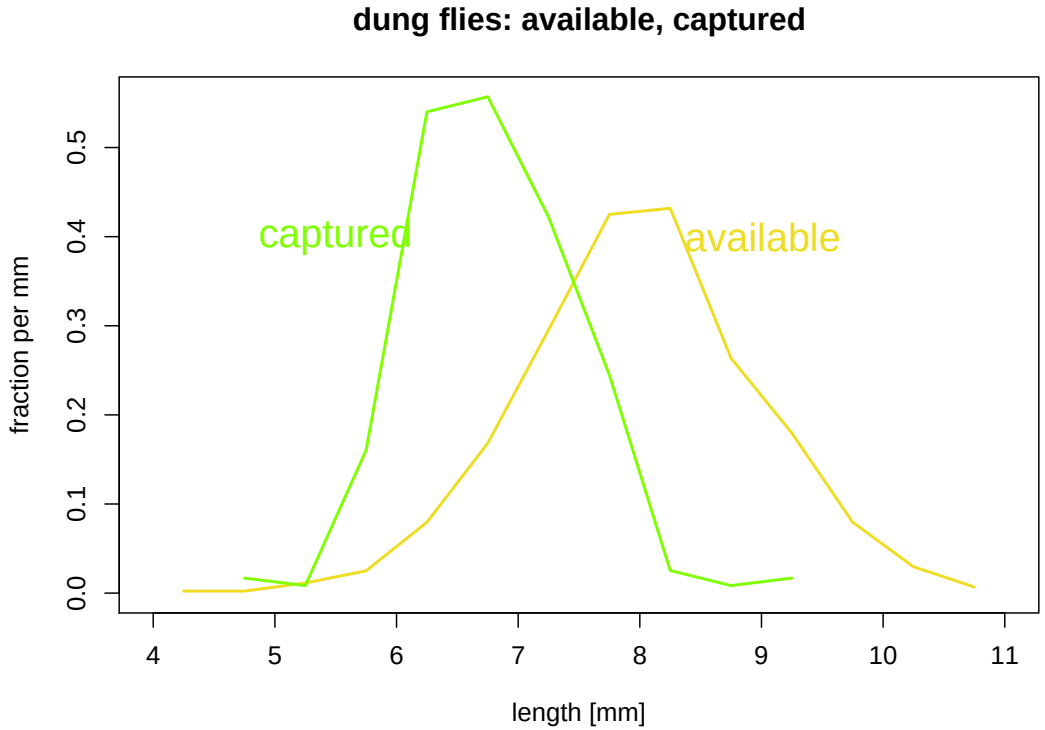

captured dung flies
mean= 6.79
number
0
10
20
30
40
50
60
4
5
6
7
8
9
10
11
length [mm]


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]

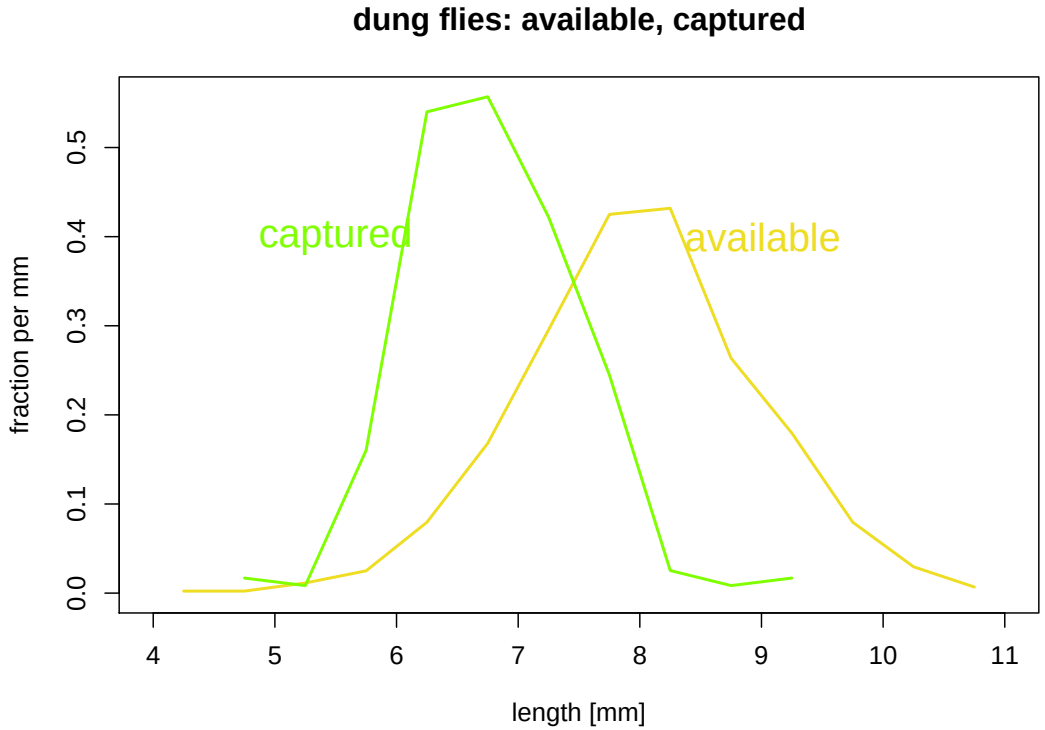
Vergleich der Größenverteilungen

	captured	available
Mittelwert		



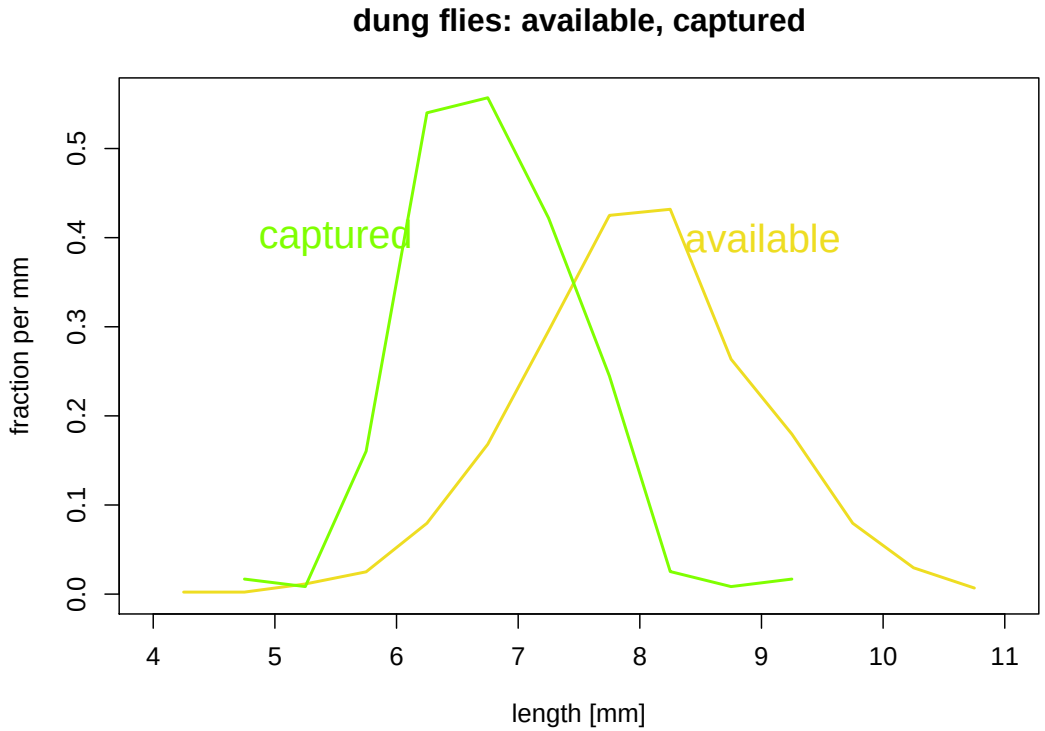
Vergleich der Größenverteilungen

	captured		available
Mittelwert		<	



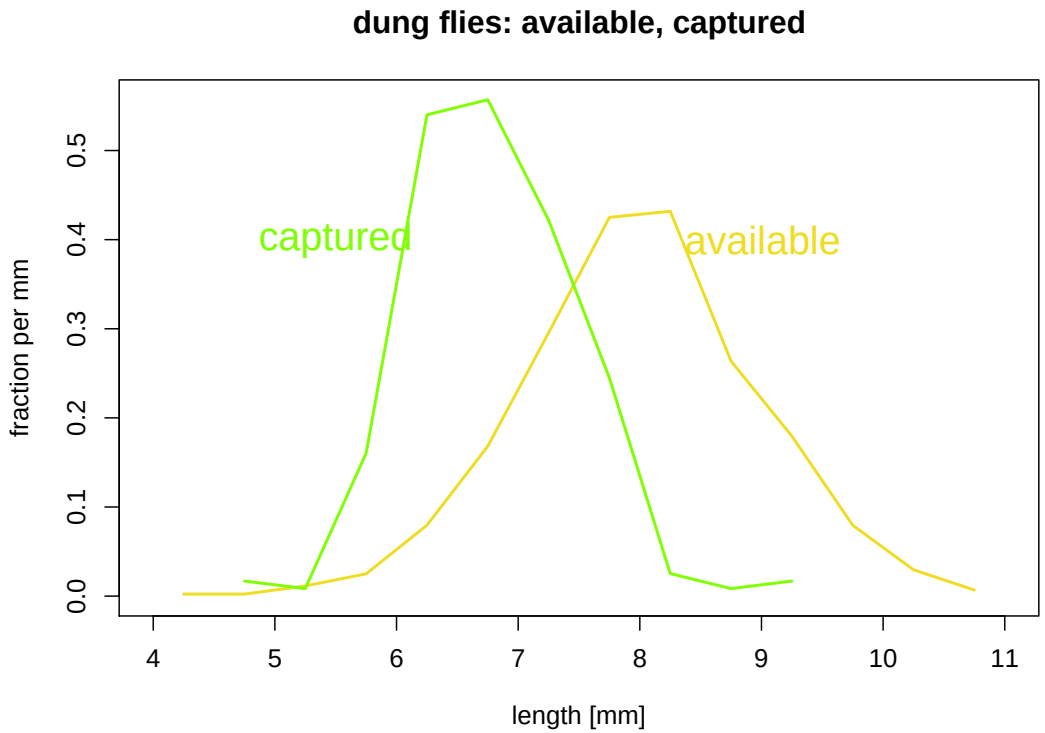
Vergleich der Größenverteilungen

	captured		available
Mittelwert	6.29	$<$	7.99



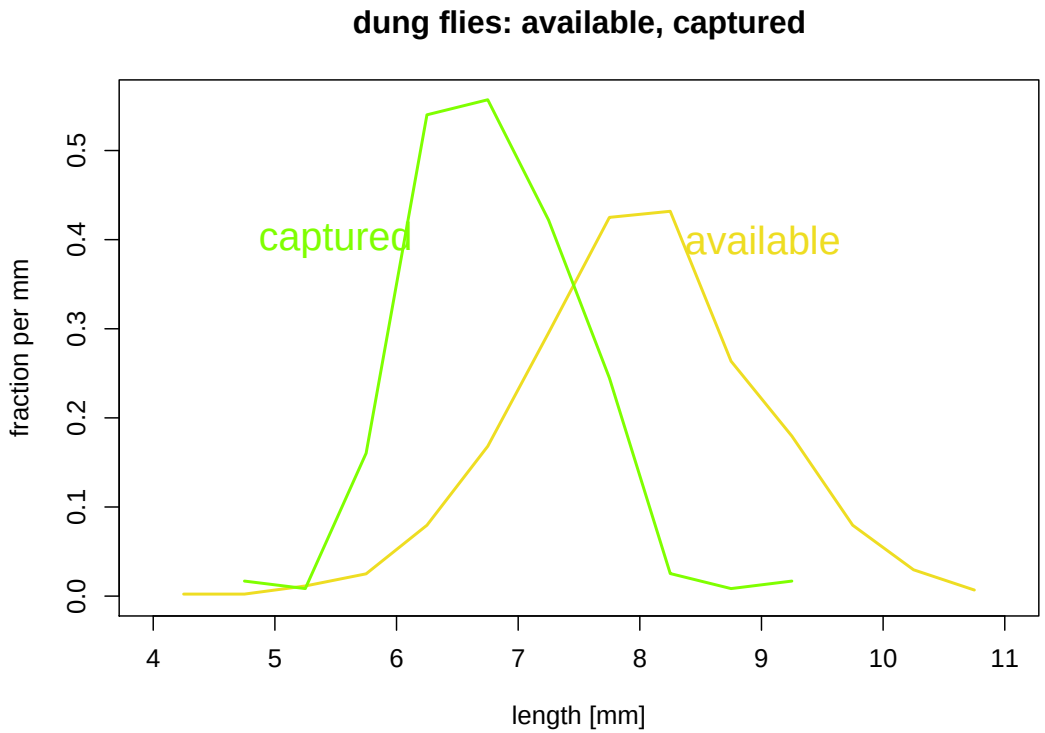
Vergleich der Größenverteilungen

	captured		available
Mittelwert	6.29	$<$	7.99
Standardabweichung			



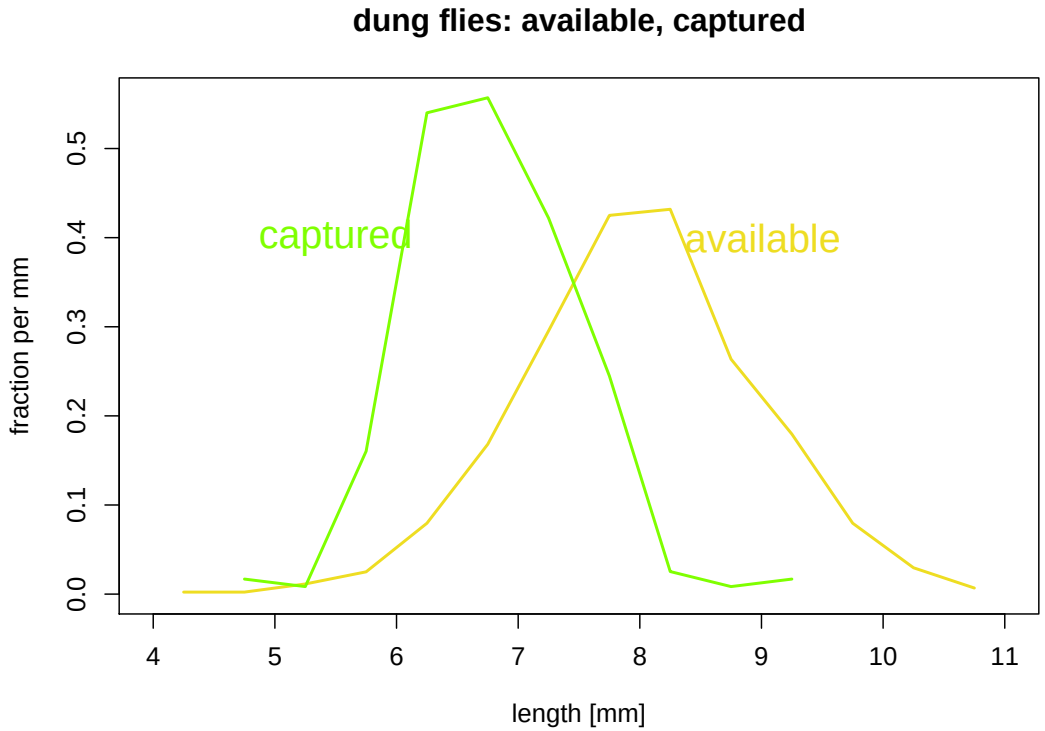
Vergleich der Größenverteilungen

	captured		available
Mittelwert	6.29	$<$	7.99
Standardabweichung		$<$	



Vergleich der Größenverteilungen

	captured		available
Mittelwert	6.29	$<$	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.

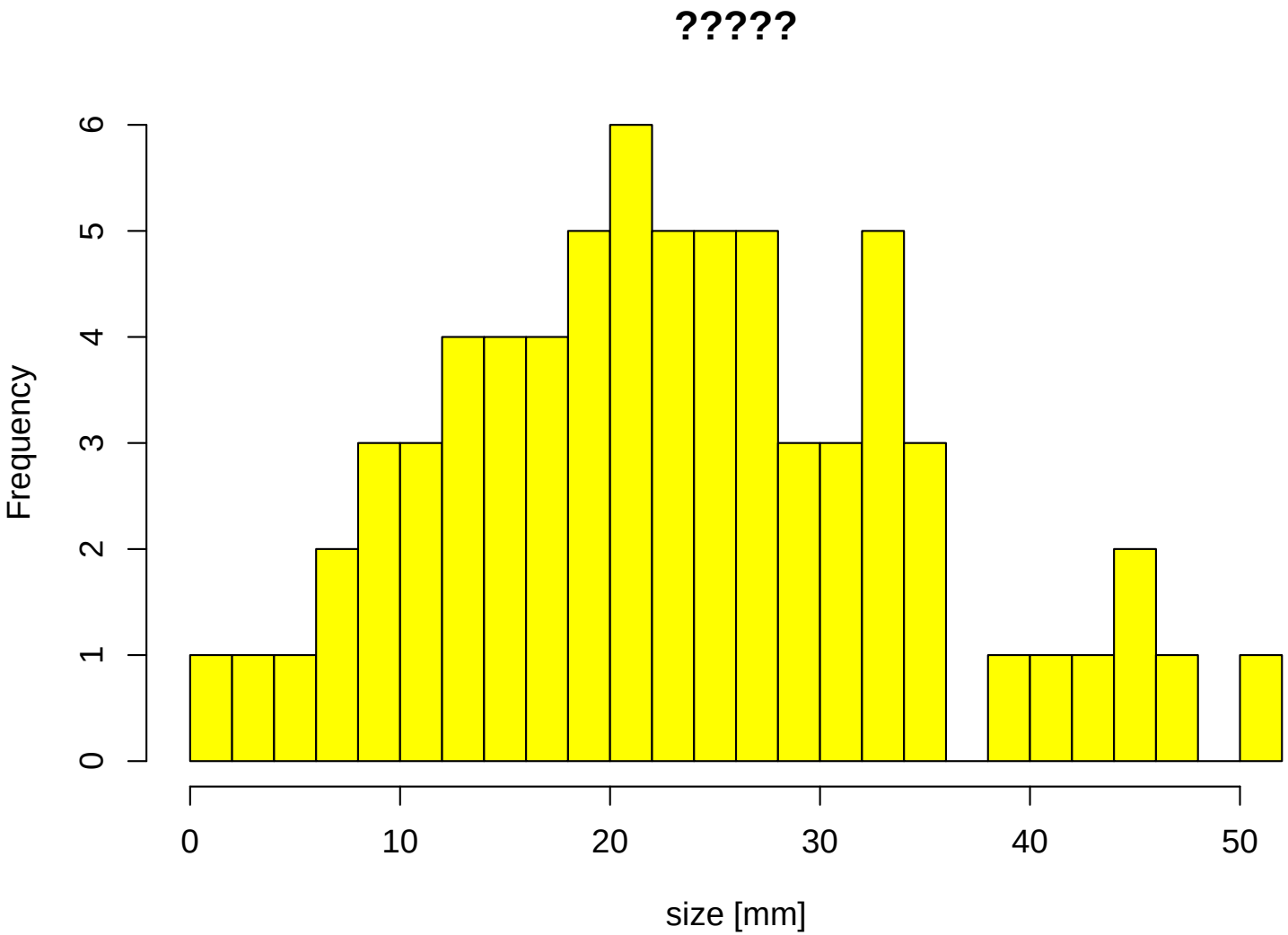
Inhalt

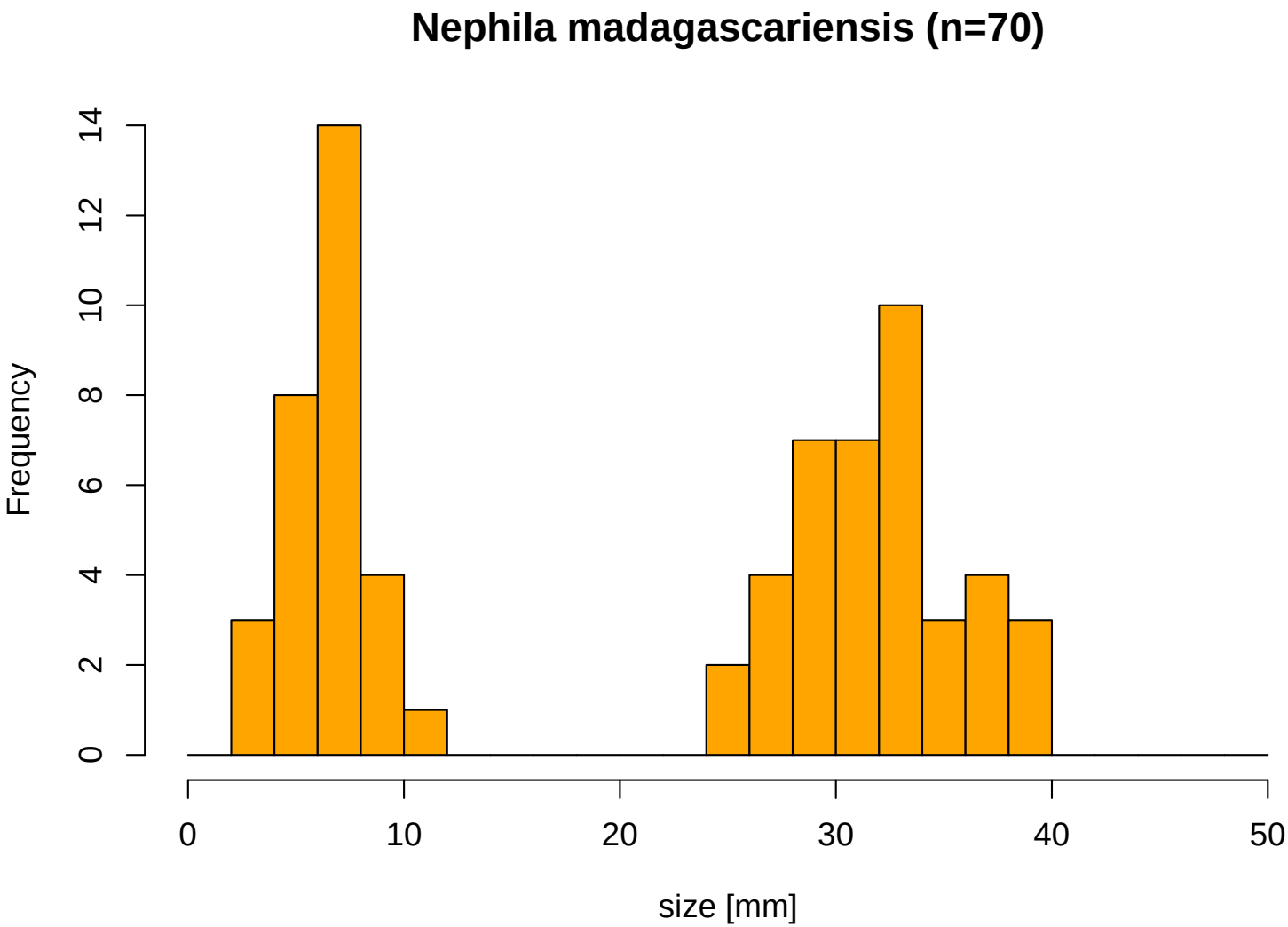
Nephila madagascariensis

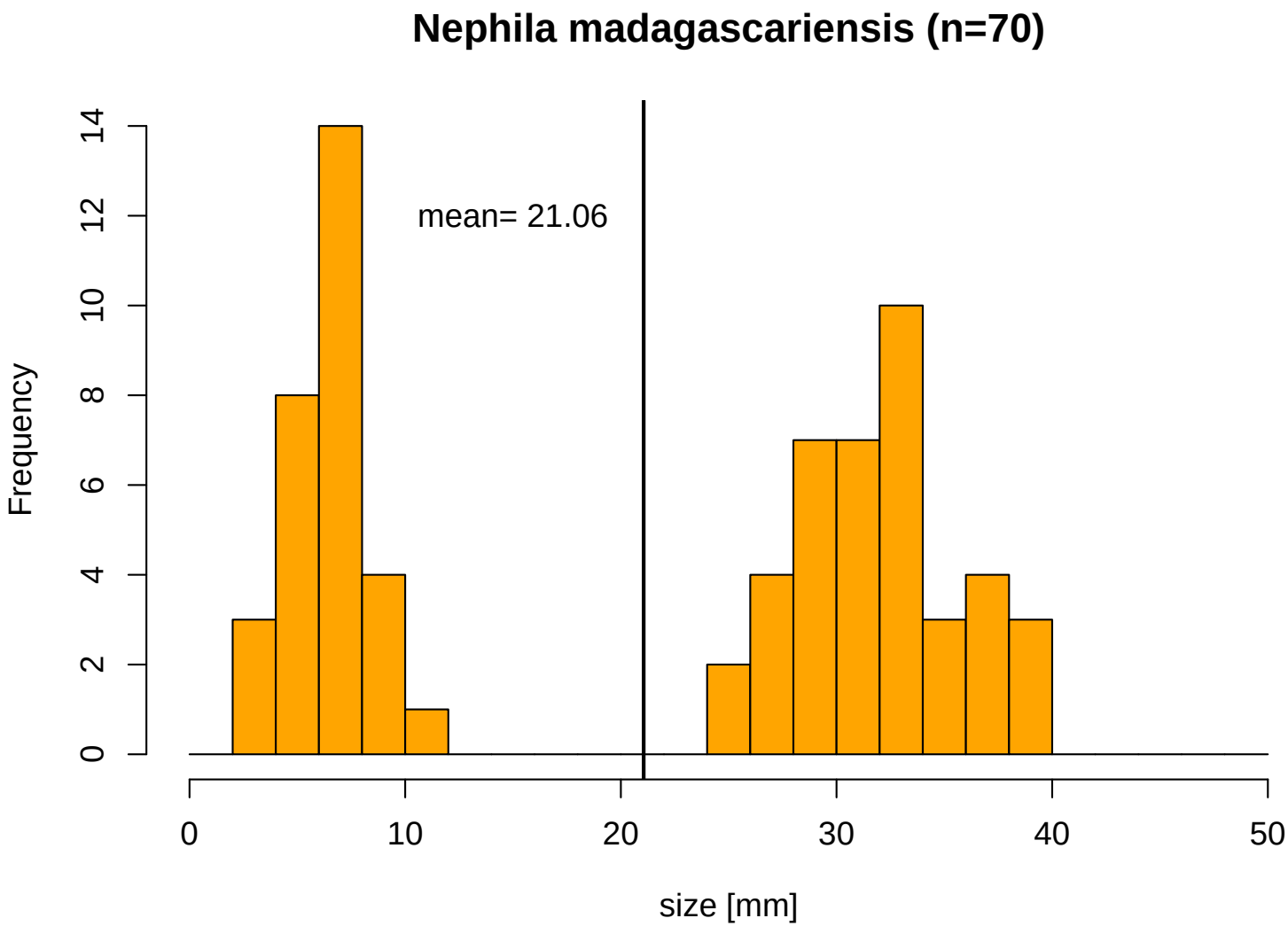
image (c) by Bernard Gagnon

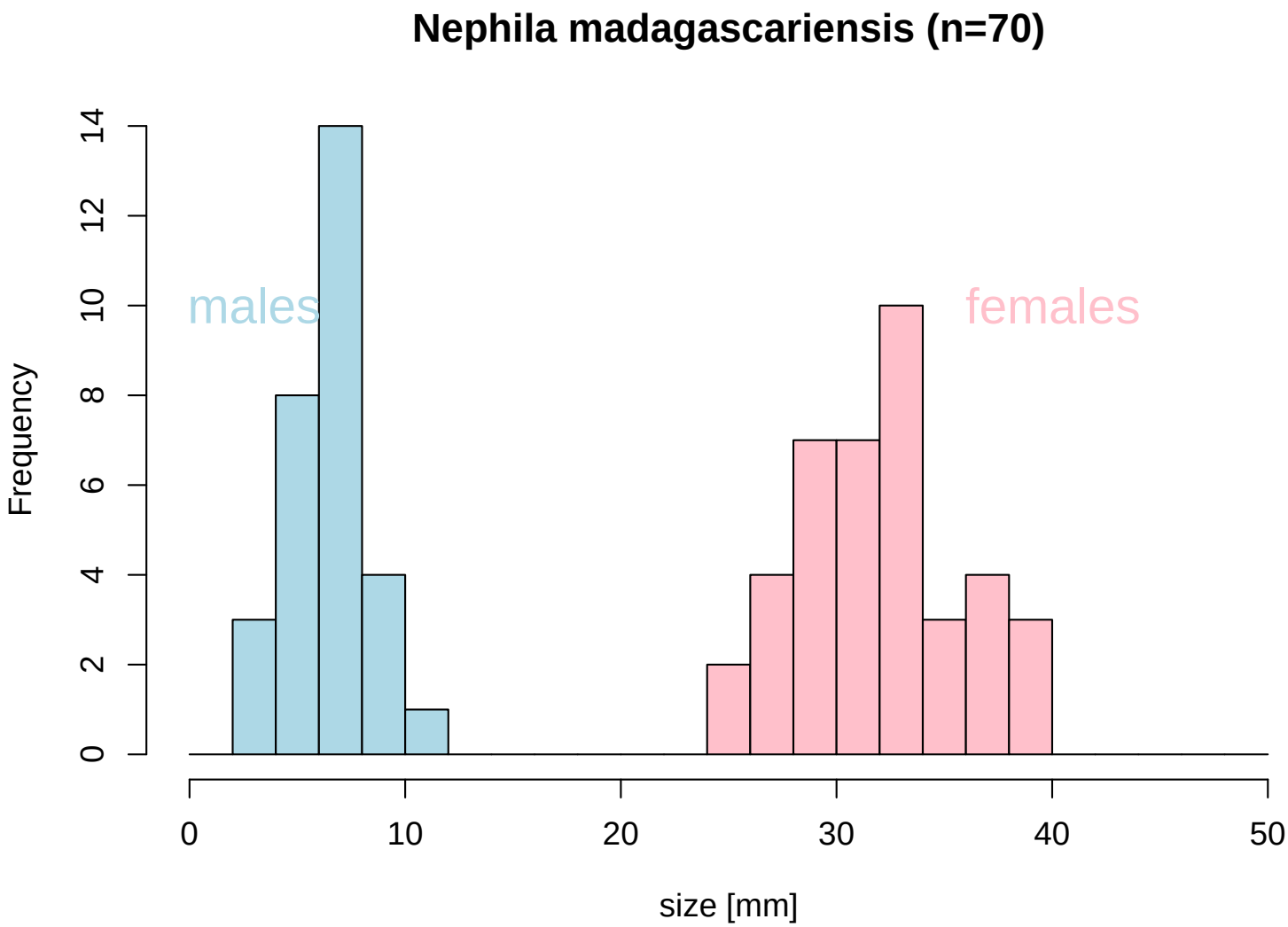
Simulierte Daten:

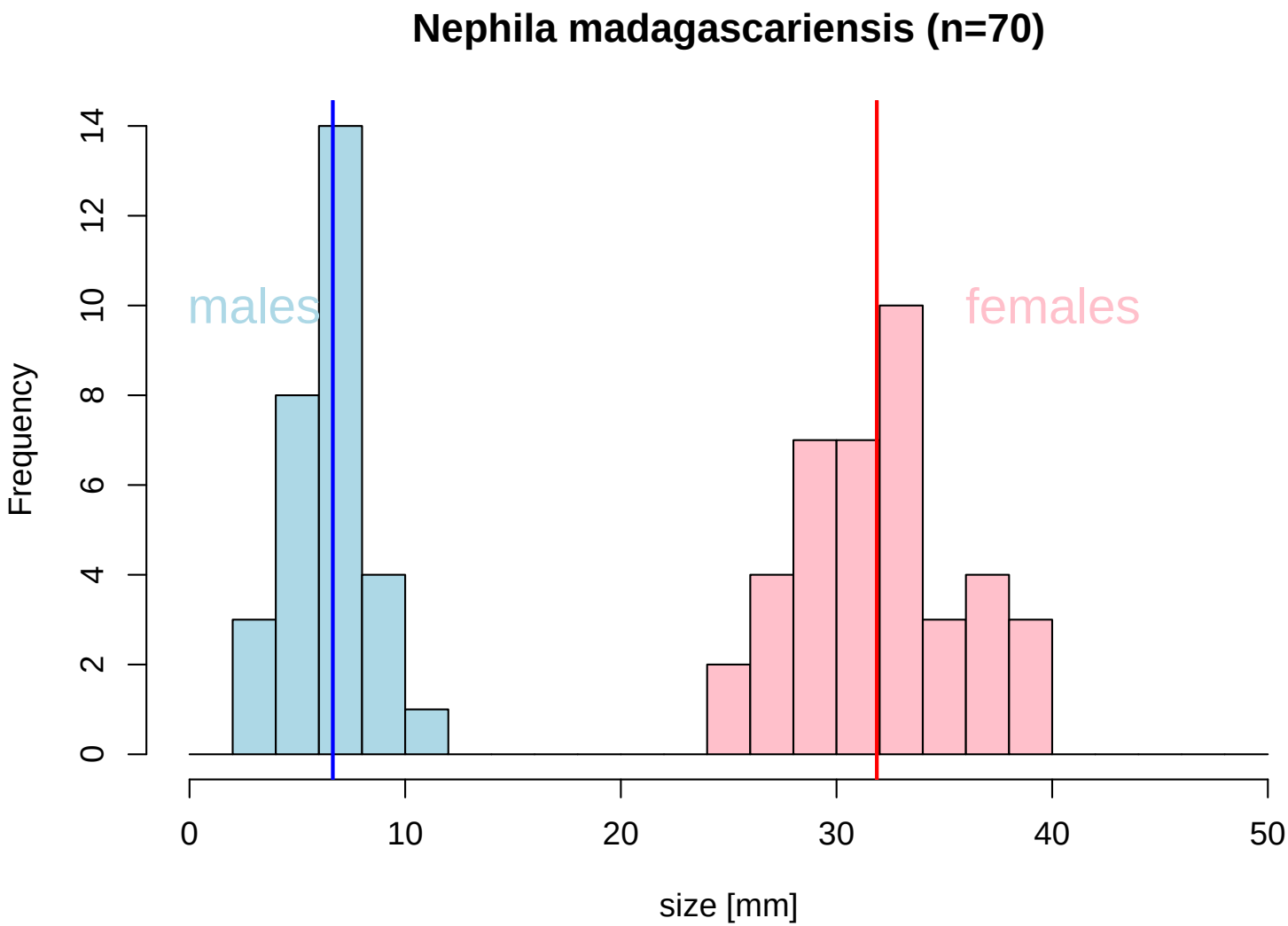
Eine Stichprobe von 70 Spinnen
Mittlere Größe: 21,06 mm
Standardabweichung der Größe: 12,94 mm


?????
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


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.


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.

Wir verwenden hier wieder simulierte Daten, da die Originaldaten nicht zur Verfügung stehen.

Anpassung an Kupfer?

- Pflanzen, denen das Kupfer schadet, haben kürzere Wurzeln.

Anpassung an Kupfer?

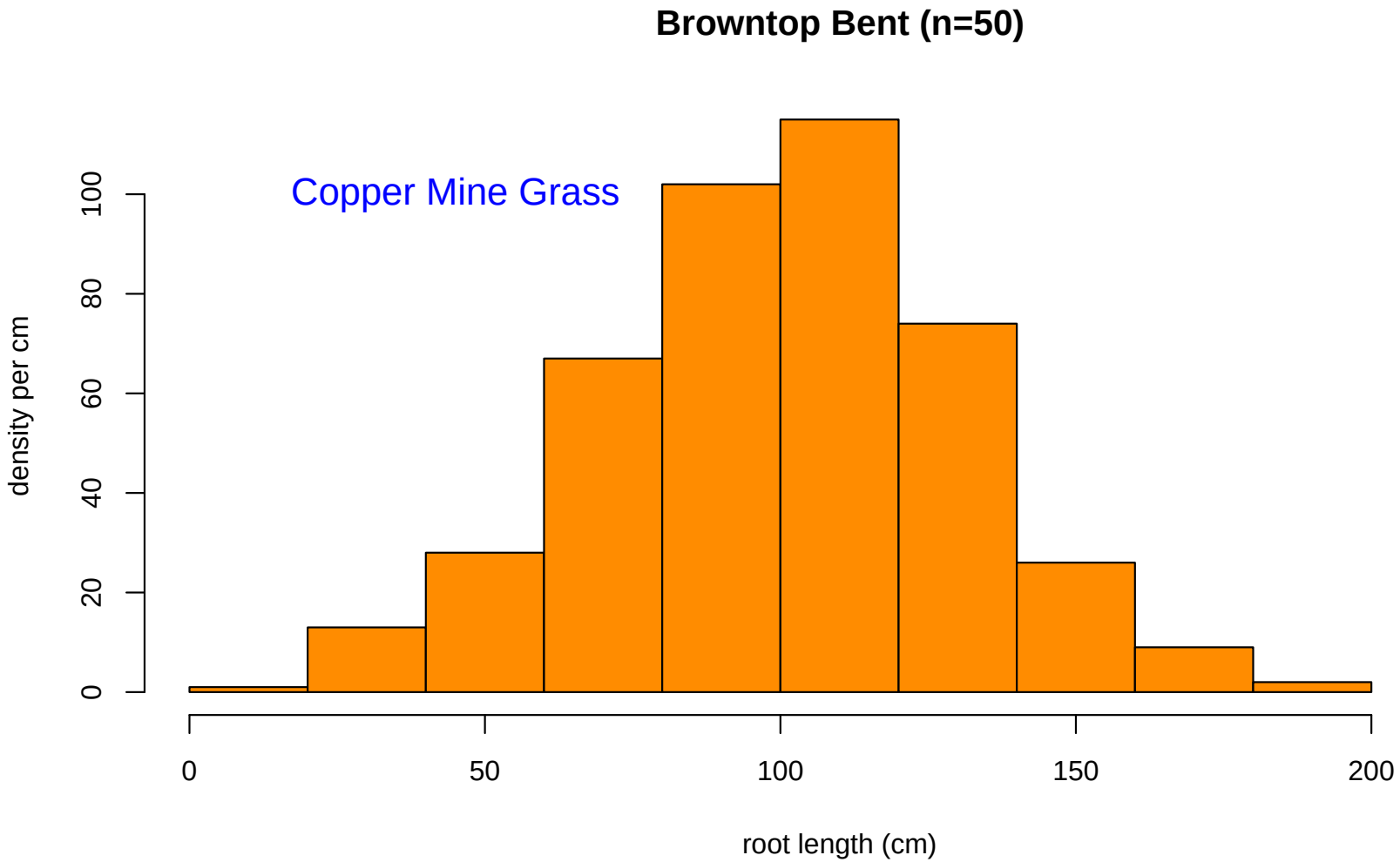
- Pflanzen, denen das Kupfer schadet, haben kürzere Wurzeln.
- Die Wurzellängen von Pflanzen aus der Umgebung von Kupferminen wird gemessen.

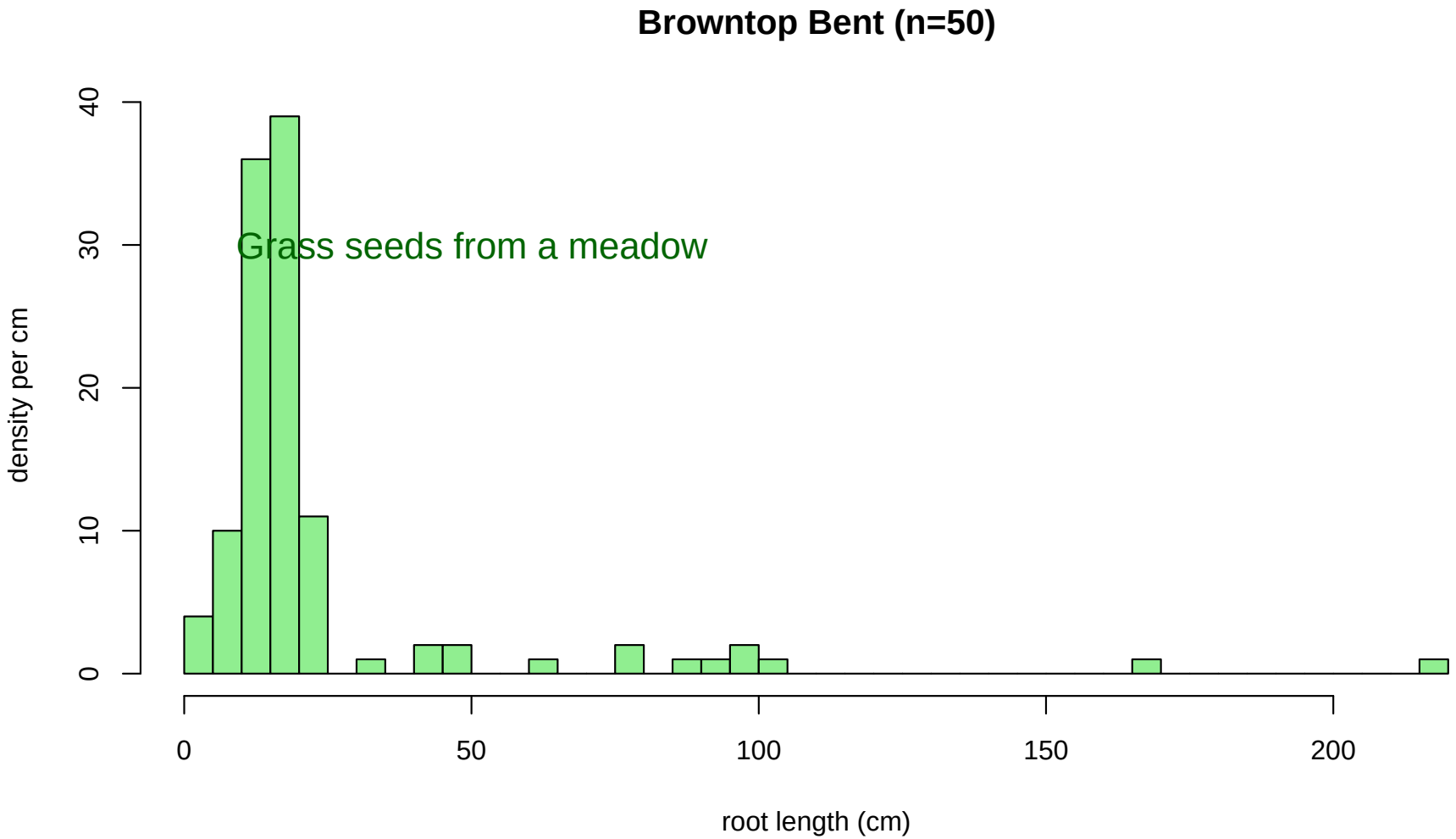
Anpassung an Kupfer?

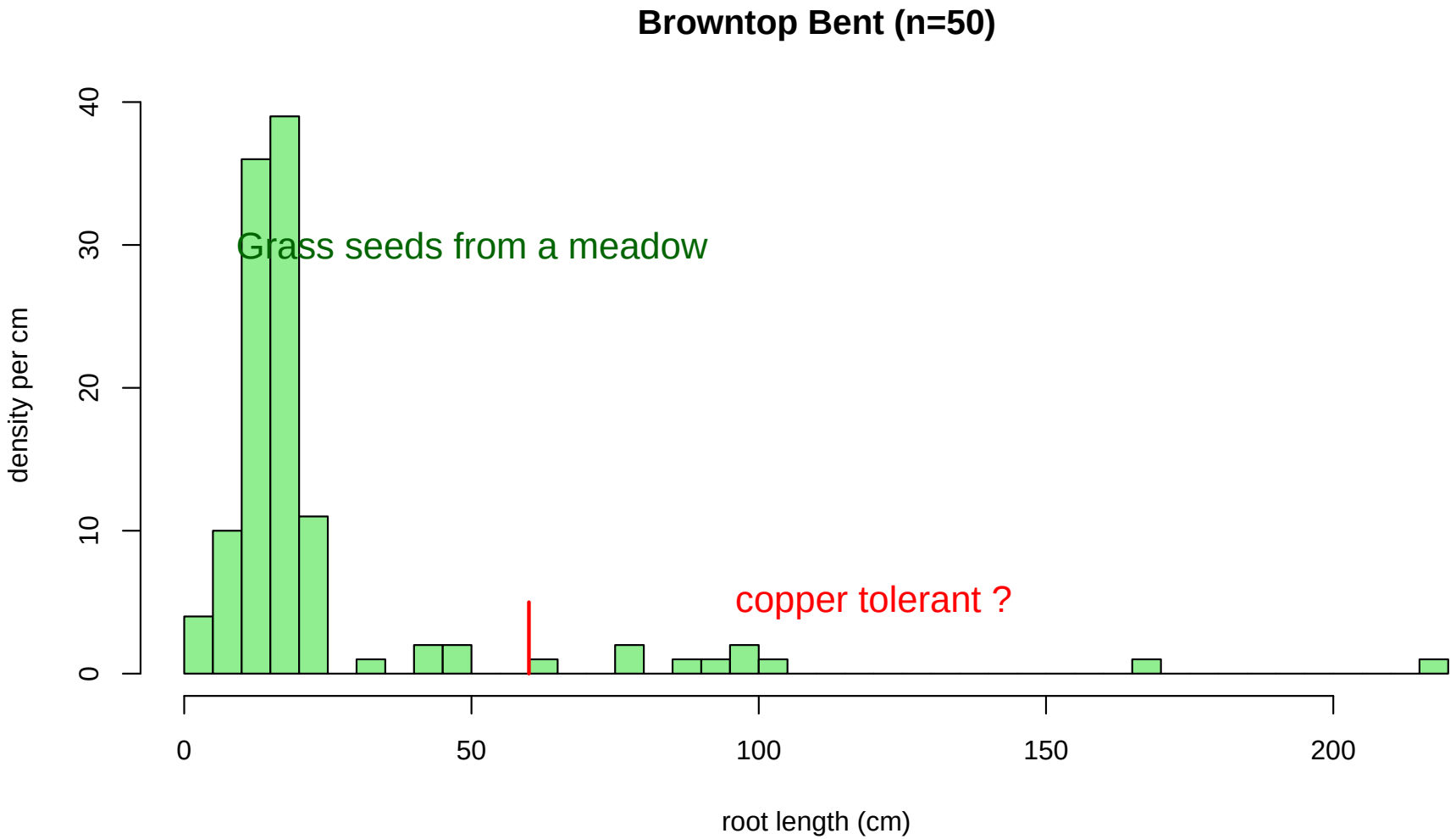
- Pflanzen, denen das Kupfer schadet, haben kürzere Wurzeln.
- Die Wurzellängen von Pflanzen aus der Umgebung von Kupferminen wird gemessen.
- Samen von unbelasteten Wiesen werden bei Kupferminen eingesät.

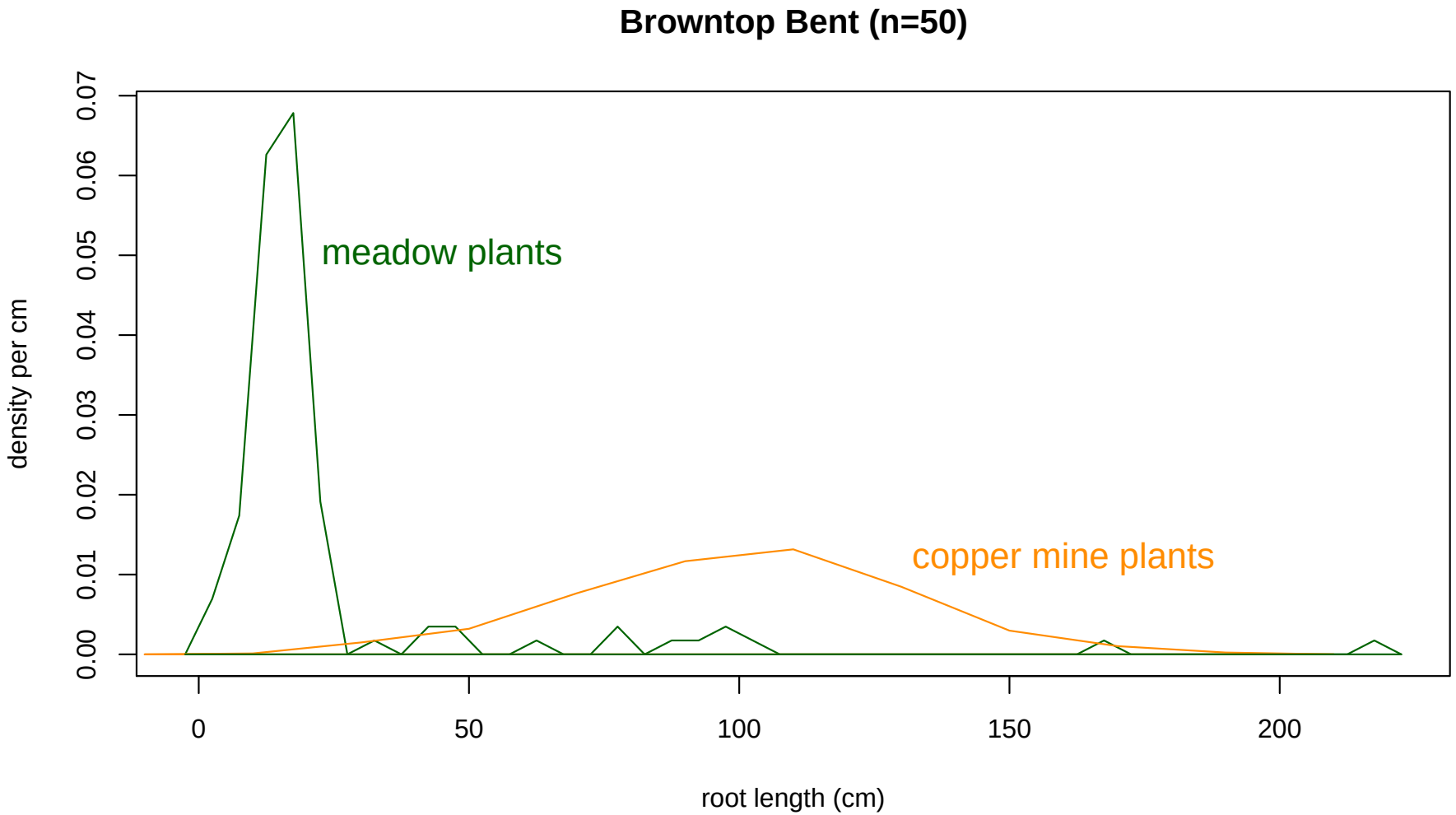
Anpassung an Kupfer?

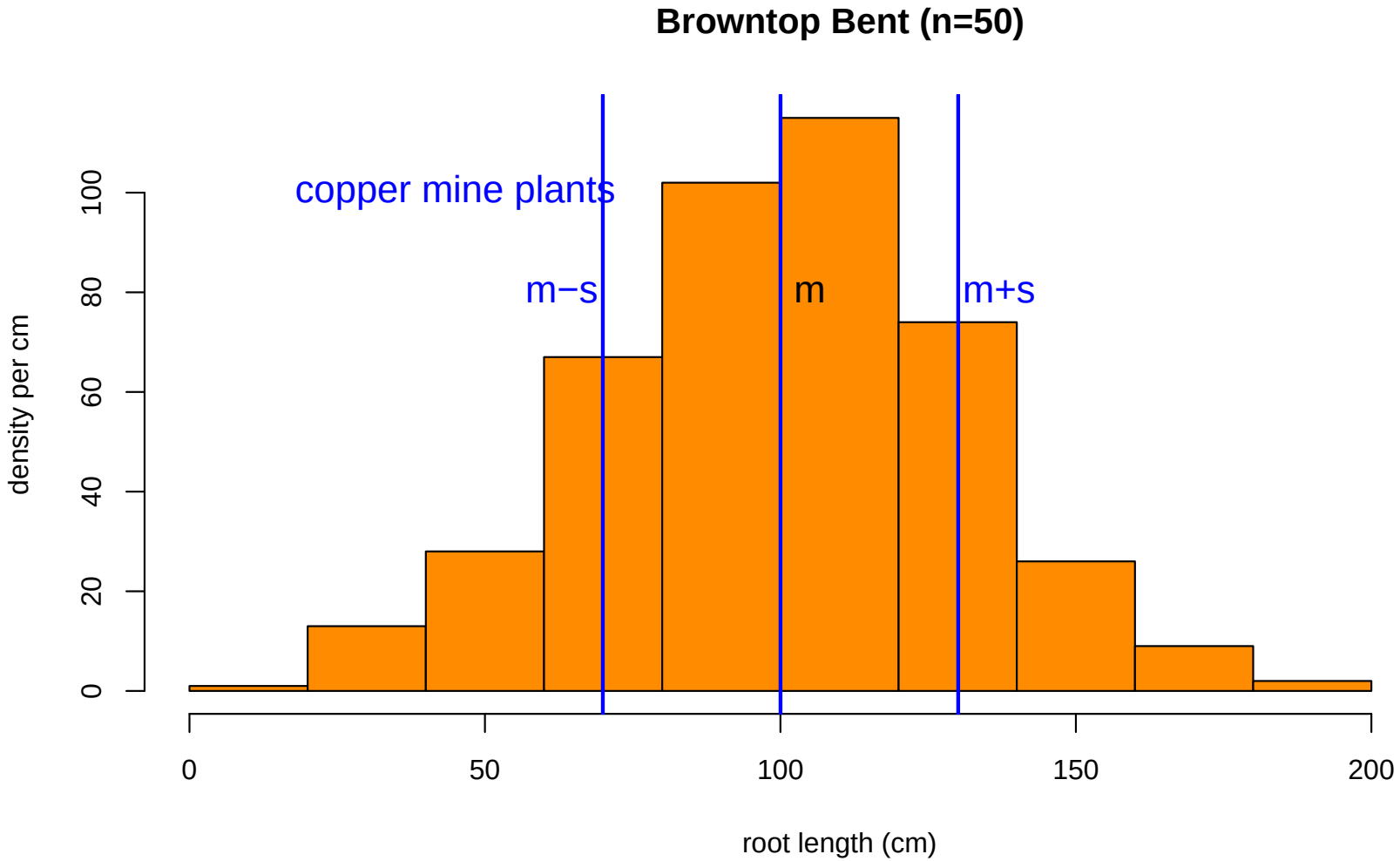
- Pflanzen, denen das Kupfer schadet, haben kürzere Wurzeln.
- Die Wurzellängen von Pflanzen aus der Umgebung von Kupferminen wird gemessen.
- Samen von unbelasteten Wiesen werden bei Kupferminen eingesät.
- Die Wurzellängen dieser “Wiesenpflanzen” werden gemessen.

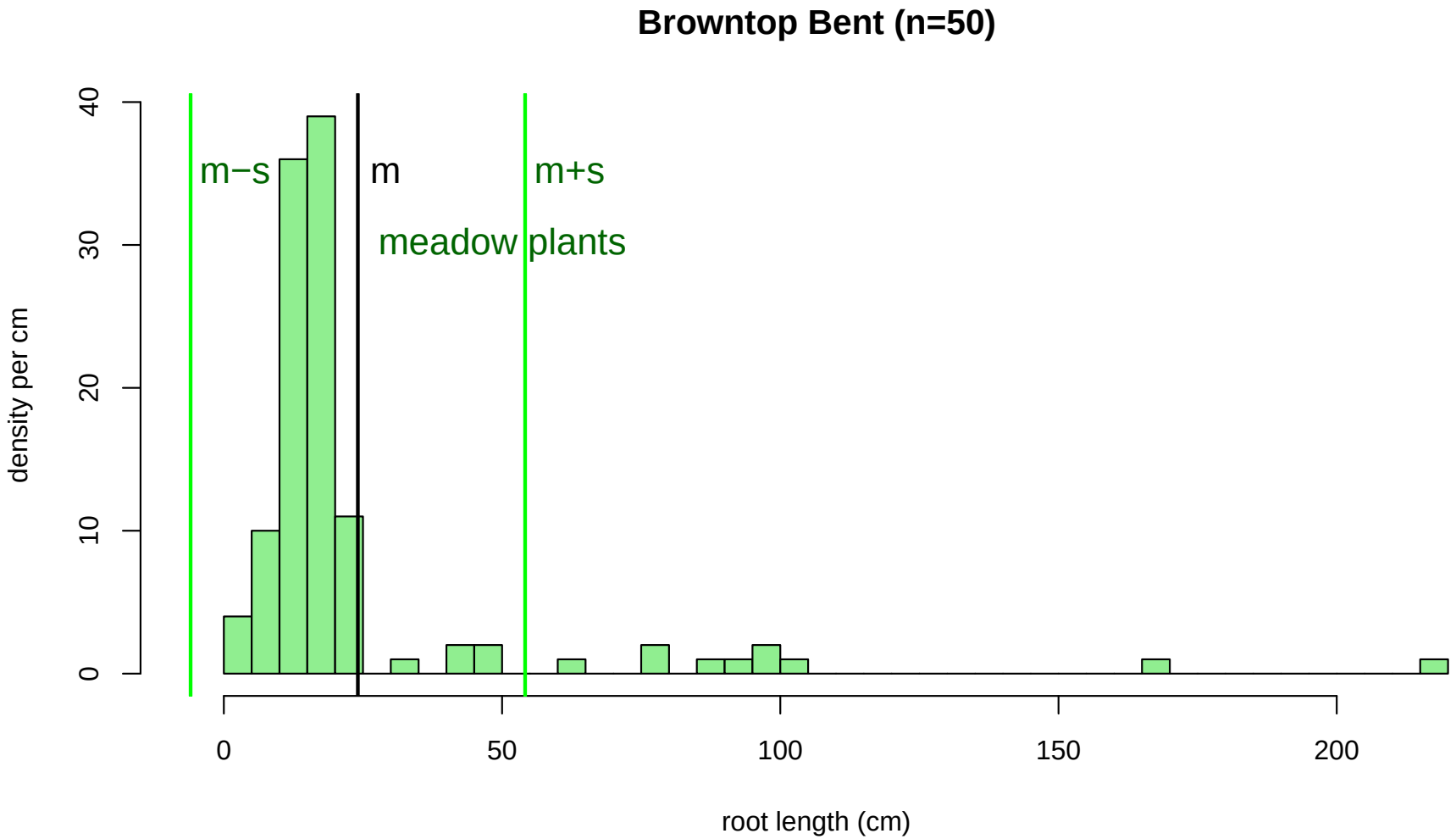

Browntop Bent (n=50)
Copper Mine Grass
density per cm
0
20
40
60
80
100
0
50
100
150
200
root length (cm)


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)
density per cm
0.00
0.01
0.02
0.03
0.04
0.05
0.06
0.07
meadow plants
copper mine plants
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!

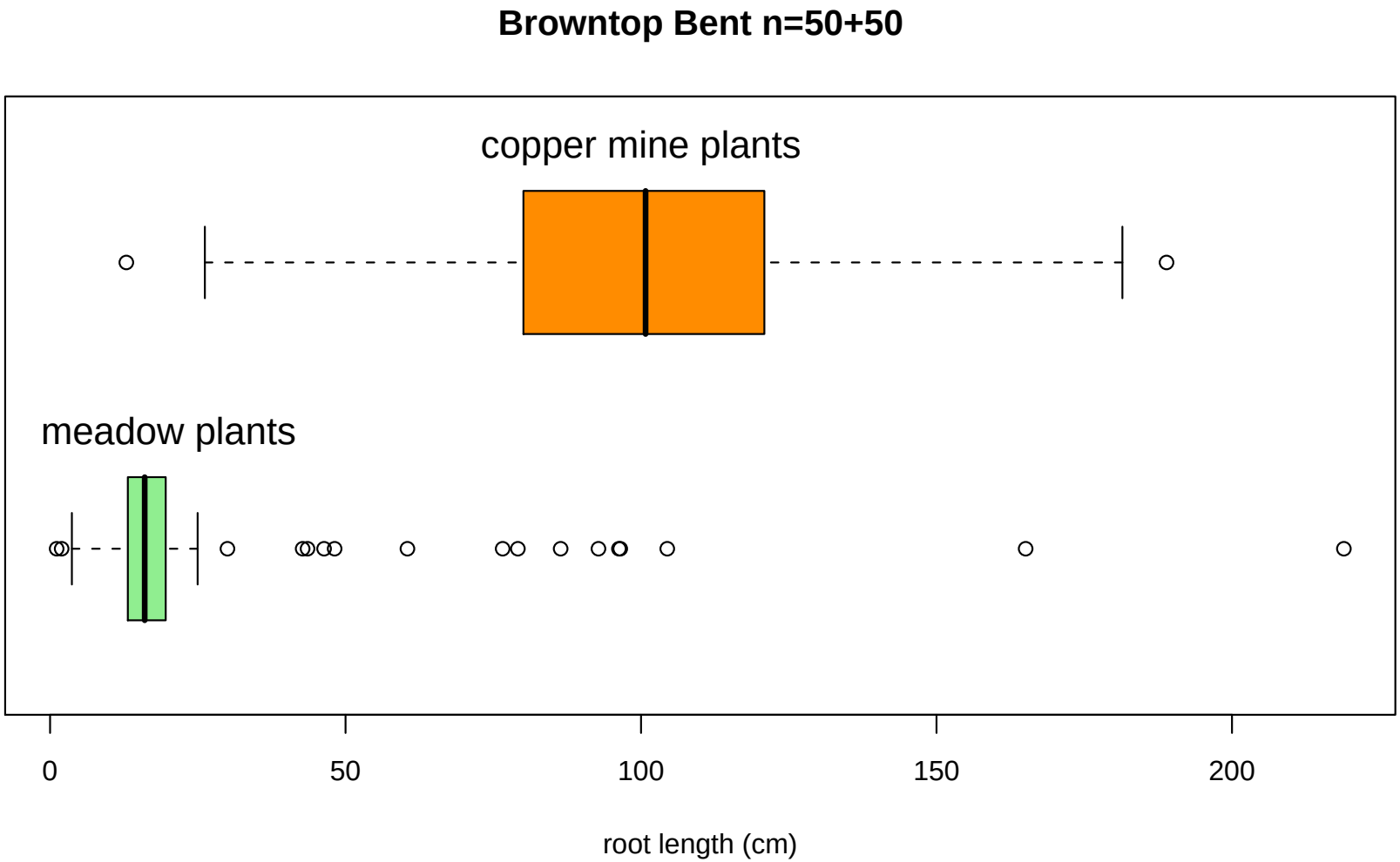
Fazit des Straußgras-Beispiels

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.

Inhalt

Was ist R?

- Man kann R als (sehr mächtigen) „Statistik-Taschenrechner“ verwenden.

Was ist R?

- Man kann R als (sehr mächtigen) „Statistik-Taschenrechner“ verwenden.

  (R ist eine für die Statistik und für stochastische Simulation entwickelte Programmiersprache, zudem sind viele statistische Standardverfahren bereits in R implementiert oder als Zusatzpaket verfügbar.

  R wird auch im Stochastik-Praktikums verwendet.)

Was ist R?

- Man kann R als (sehr mächtigen) „Statistik-Taschenrechner“ verwenden.

  (R ist eine für die Statistik und für stochastische Simulation entwickelte Programmiersprache, zudem sind viele statistische Standardverfahren bereits in R implementiert oder als Zusatzpaket verfügbar.

  R wird auch im Stochastik-Praktikums verwendet.)
- R hat eine sehr aktive Benutzer- und Entwicklergemeinde (die nahezu alle Bereiche der Statistik und viele Anwendungsbereiche (z.B. Populationsgenetik, Finanzmathematik) überdeckt).

Was ist R?

- Man kann R als (sehr mächtigen) „Statistik-Taschenrechner“ verwenden.

  (R ist eine für die Statistik und für stochastische Simulation entwickelte Programmiersprache, zudem sind viele statistische Standardverfahren bereits in R implementiert oder als Zusatzpaket verfügbar.
  R wird auch im Stochastik-Praktikums verwendet.)
- R hat eine sehr aktive Benutzer- und Entwicklergemeinde (die nahezu alle Bereiche der Statistik und viele Anwendungsbereiche (z.B. Populationsgenetik, Finanzmathematik) überdeckt).
- R ist frei verfügbar unter der GNU general public license, für (nahezu) alle Rechnerarchitekturen erhältlich:
  http://www.r-project.org/
- R ist auf ZDV-Rechnern installiert.

R installieren, starten, anhalten

Installation: Windows, Mac OS: Binaries von
http://www.r-project.org/ (siehe Link Download, Packages, CRAN dort)
Linux: Für die meisten Distributionen gibt es fertige Pakete

Fragen oder Probleme:
Tutorium (J. Blauth), Do 10-12, Raum MI 2

R starten: Windows, Mac OS: Icon (ggf. aus Menu) anklicken,
Linux/Unix: > R auf einer Konsole
(oder mit ESS aus Emacs heraus, oder aus rstudio, ...).

R beenden: q() (fragt, ob Daten gespeichert werden sollen)

laufende Rechnungen unterbrechen: CTRL-C

Inhalt

Datensatz `x` in R eingeben

```
x <- c(  53,52,41,41,42,58,40,43,42,38,43,49,34,51,45,
         39,41,45,45,39,37,36,42,44,47,43,46,43,43,45,
         42,52,49,44,50,40,47,46,50,50,41,51,41,47,42,
         52,36,46,42,56,39,40,36,42,36,36,47,45,47,49 )
```

(aus Datei einlesen: `x <- scan('Dateiname')`)

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)
```

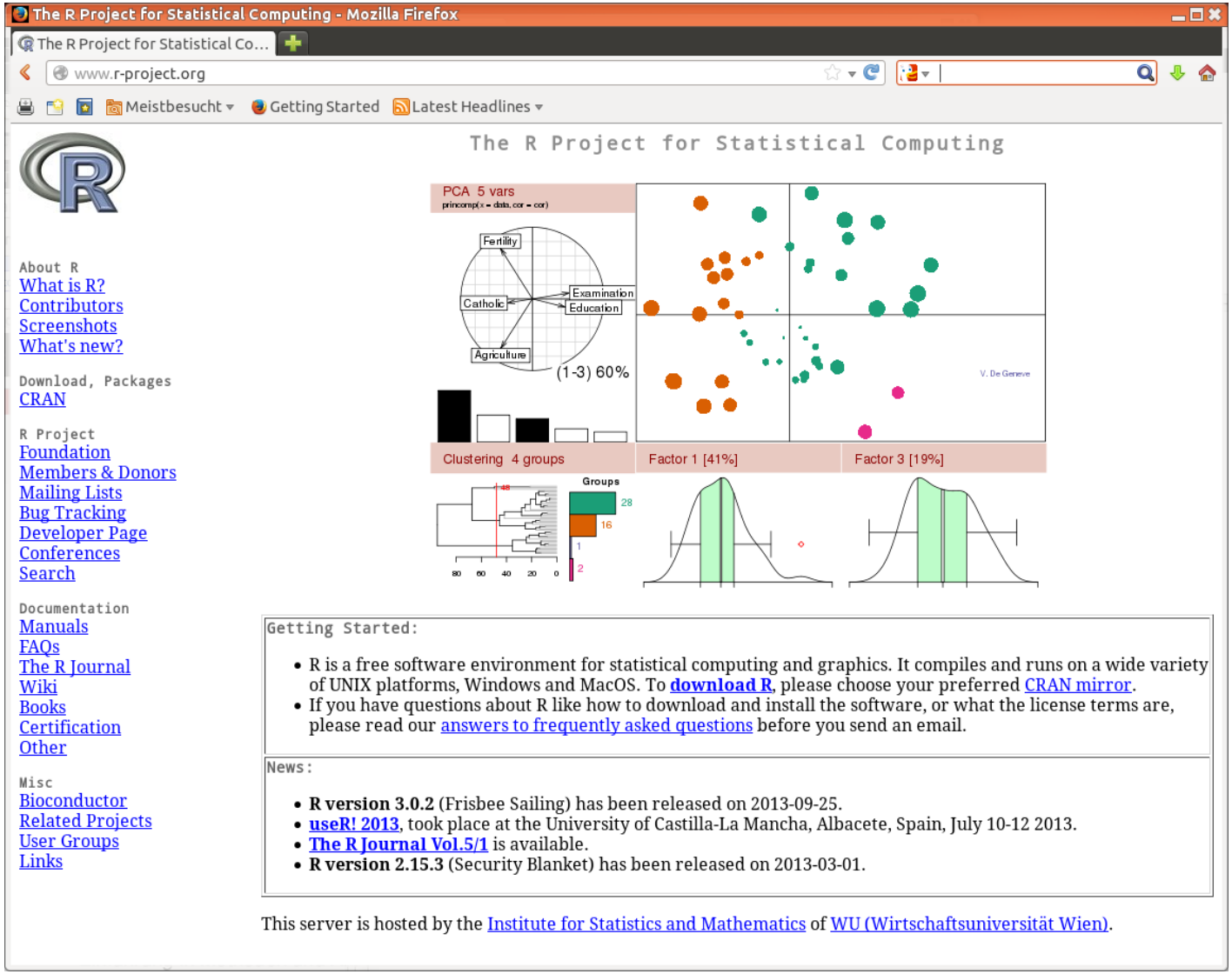
Boxplot, Histogramm

```
boxplot(x)
hist(x)
```

Nur zur Information: Literatur zu R

Wir werden im Zusammenhang der Vorlesung nur (fakultativ) einige wenige R-Befehle ansprechen, so dass Sie für etwaige freiwillige „R-Schnupperei“ i.A. keine Literatur benötigen werden.

- „Standardreferenz“: W.N. Venables et al, *An Introduction to R*, http://cran.r-project.org/manuals.html
- Günther Sawitzki, *Einführung in R*, http://sintro.r-forge.r-project.org/
- William N. Venables, Brian D. Ripley, *Modern applied statistics with S* („Standardlehrbuch“, UB Lehrbuchsammlung)
- Lothar Sachs and Jürgen Hedderich, *Angewandte Statistik – Methodensammlung mit R* (E-Book, UB)
- Christine Duller, *Einführung in die nichtparametrische Statistik mit SAS und R : ein anwendungsorientiertes Lehr- und Arbeitsbuch* (E-Book, UB)
- Helge Toutenburg, Christian Heumann, *Deskriptive Statistik : Eine Einführung in Methoden und Anwendungen mit R und SPSS* (E-Book, UB)
- Uwe Ligges, *Programmieren mit R* (E-Book, UB)


The R Project for Statistical Computing - Mozilla Firefox
The R Project for Statistical Co...
www.r-project.org
Meistbesucht
Getting Started
Latest Headlines
The R Project for Statistical Computing
About R
What is R?
Contributors
Screenshots
What's new?
Download, Packages
CRAN
R Project
Foundation
Members & Donors
Mailing Lists
Bug Tracking
Developer Page
Conferences
Search
Documentation
Manuals
FAQs
The R Journal
Wiki
Books
Certification
Other
Misc
Bioconductor
Related Projects
User Groups
Links
PCA 5 vars
Fertility
Examination
Education
Catholic
Agriculture
(1-3) 60%
V. De Geneve
Clustering 4 groups
Factor 1 [41%]
Factor 3 [19%]
Groups
Getting Started:
R is a free software environment for statistical computing and graphics. It compiles and runs on a wide variety of UNIX platforms, Windows and MacOS. To download R, please choose your preferred CRAN mirror.
If you have questions about R like how to download and install the software, or what the license terms are, please read our answers to frequently asked questions before you send an email.
News:
R version 3.0.2 (Frisbee Sailing) has been released on 2013-09-25.
useR! 2013, took place at the University of Castilla-La Mancha, Albacete, Spain, July 10-12 2013.
The R Journal Vol.5/1 is available.
R version 2.15.3 (Security Blanket) has been released on 2013-03-01.
This server is hosted by the Institute for Statistics and Mathematics of WU (Wirtschaftsuniversität Wien).

Nochmal: Idee der Statistik

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

Nochmal: 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.

Nochmal: 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.