Biostatistik, WS 2010/2011

Deskriptive Statistik

Matthias Birkner

http://www.mathematik.uni-mainz.de/~birkner/Biostatistik1011/

19.11.2010

JOHANNES GUTENBERG UNIVERSITÄT MAINZ

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

Andrejs Dunkels

Was ist Statistik?

Die Natur ist voller Variabilität.

Wie geht man mit variablen Daten um?

Es gibt eine mathematische Theorie des Zufalls:

die Stochastik.

IDEE DER STATISTIK

Variabilität

(Erscheinung der Natur)

durch

Zufall

(mathematische Abstraktion)

modellieren.

Statistik

=

Datenanalyse

mit Hilfe

stochastischer Modelle

Beispiel

Daten aus einer Diplomarbeit aus 2001 am Forschungsinstitut Senckenberg, Frankfurt am Main

Crustaceensektion

Leitung: Dr. Michael Türkay

Charybdis acutidens TÜRKAY 1985

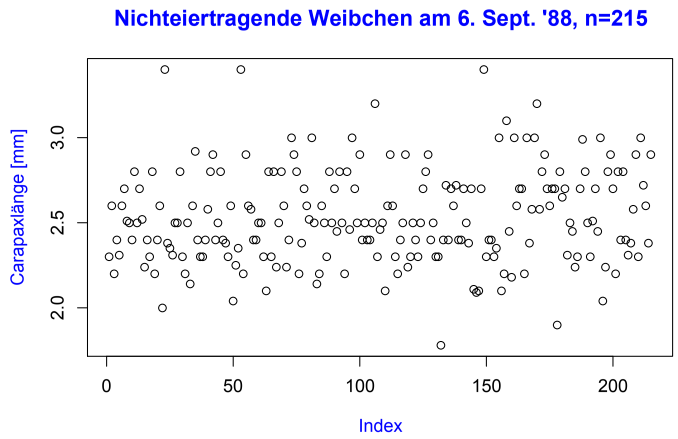
Der Springkrebs

Galathea intermedia

Helgoländer Tiefe Rinne,
Fang vom 6.9.1988

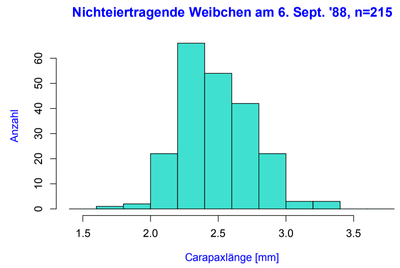
Carapaxlänge (mm):
Nichteiertragende Weibchen ($n = 215$)

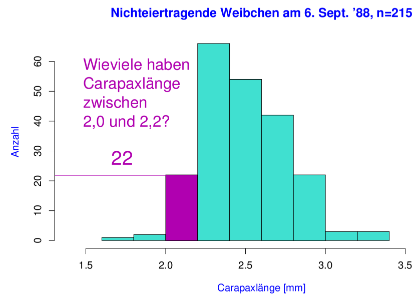
2,9	3,0	2,9	2,5	2,7	2,9	2,9	3,0
3,0	2,9	3,4	2,8	2,9	2,8	2,8	2,4
2,8	2,5	2,7	3,0	2,9	3,2	3,1	3,0
2,7	2,5	3,0	2,8	2,8	2,8	2,7	3,0
2,6	3,0	2,9	2,8	2,9	2,9	2,3	2,7
2,6	2,7	2,5	.	.	.	.	.



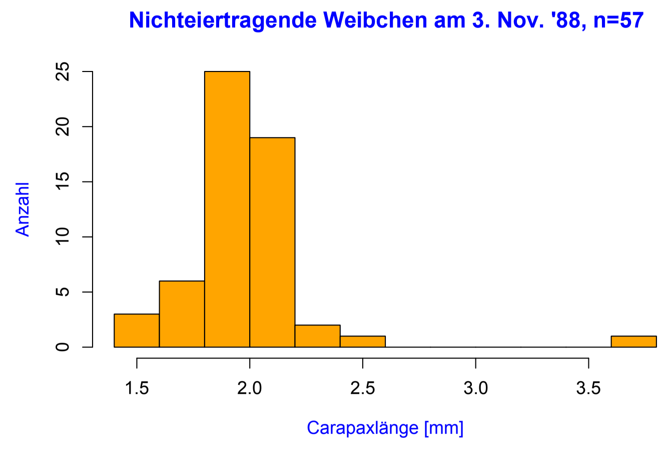
Eine Möglichkeit der graphischen Darstellung:

das Histogramm

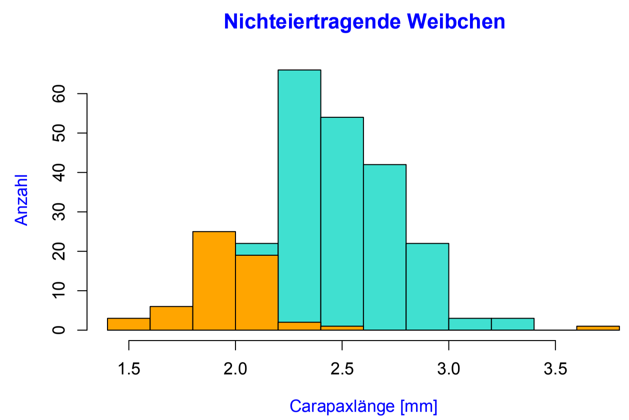




Analoge Daten zwei Monate später (3.11.88):



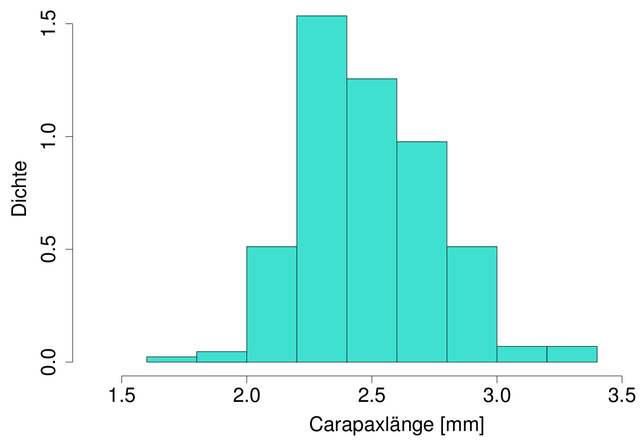
Vergleich der beiden Verteilungen



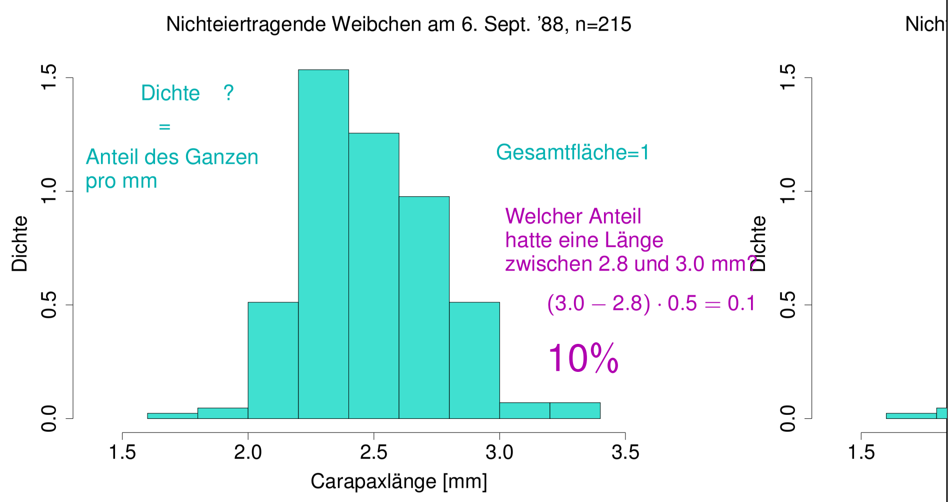
Problem: ungleiche Stichprobenumfänge: 6.Sept: $n = 215$
3.Nov : $n = 57$

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

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



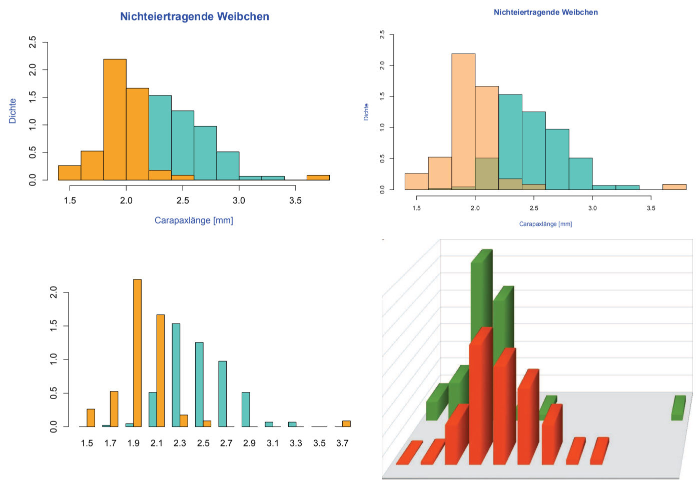
Die neue
vertikale Koordinate
ist jetzt eine
Dichte
(engl. density).



Die beiden Histogramme sind jetzt vergleichbar

(sie haben dieselbe Gesamtfläche).

Versuche, die Histogramme zusammen zu zeigen:



Ratschlag

Wenn Sie Schauwerbegestalter(in) sind:

Beeindrucken Sie Jung und Alt mit total abgefahrenen 3D-Plots!

Wenn Sie Wissenschaftler(in) werden wollen:

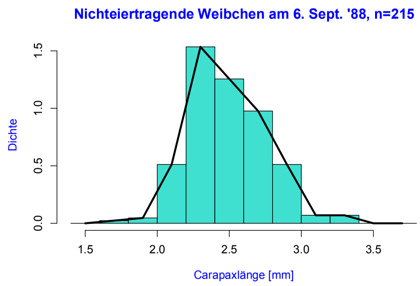
Bevorzugen Sie
einfache und klare 2D-Darstellungen.

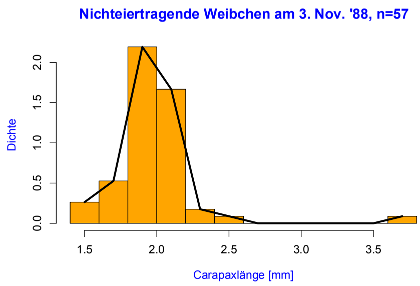
Problem

Histogramme kann man nicht ohne weiteres in demselben Graphen darstellen,

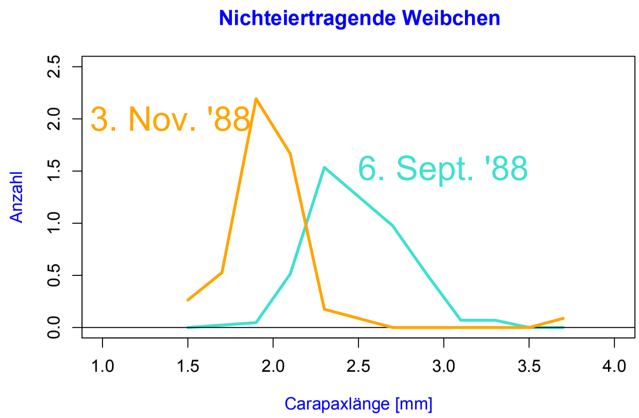
weil sie einander überdecken würden.

Einfache und klare Lösung: Dichtepolygone

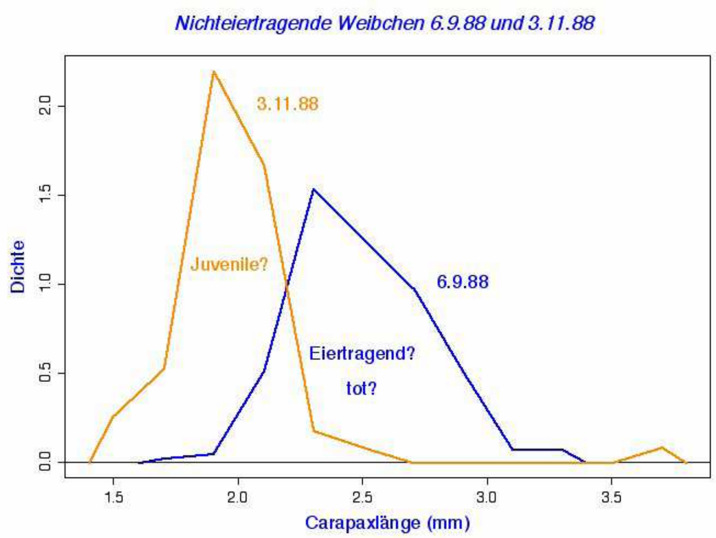




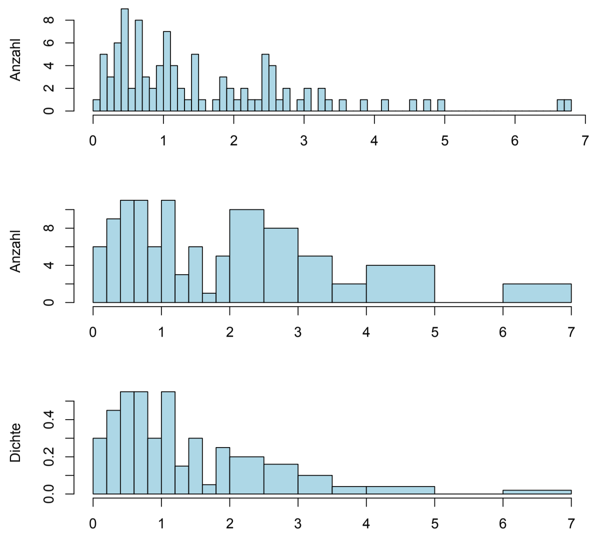
Zwei und mehr Dichtepolygone in einem Plot



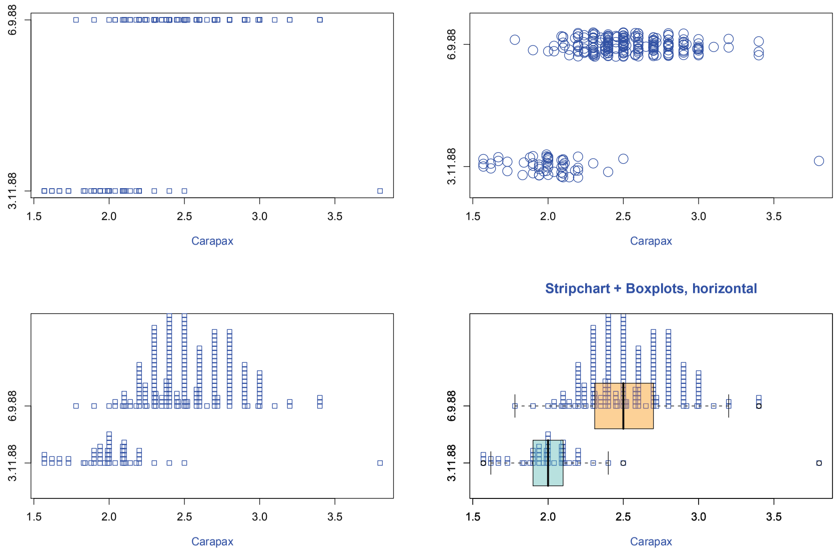
Biologische Interpretation der Verschiebung?



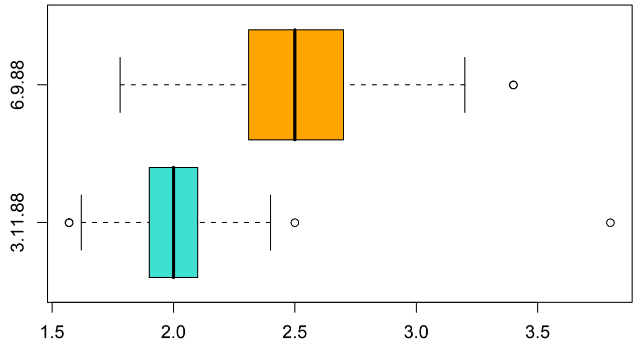
Anzahl vs. Dichte



Also: Bei Histogrammen mit ungleichmäßiger Unterteilung immer Dichten verwenden!



Boxplots, horizontal



Histogramme und Dichtepolygone
geben
ein ausführliches Bild
eines Datensatzes.

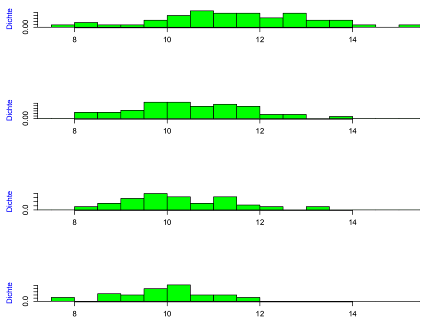
Manchmal zu ausführlich.

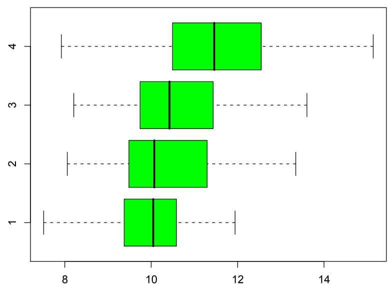
Zu viel Information erschwert den Überblick

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

Wald?

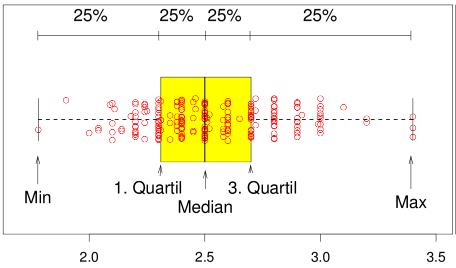
Beispiel:
Vergleich von mehreren Gruppen



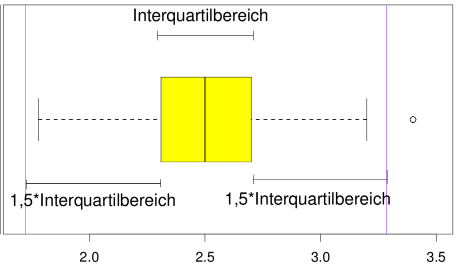


Der Boxplot

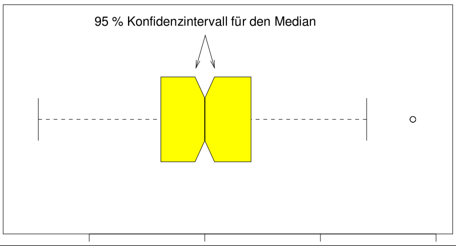
Boxplot, einfache Ausführung



Boxplot, Standardausführung



Boxplot, Profiausstattung



Beispiel:

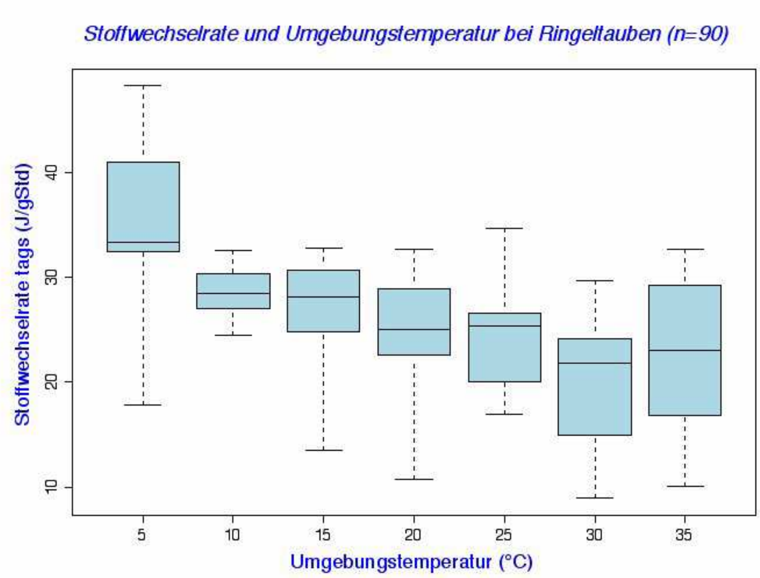
Die Ringeltaube

Palumbus palumbus

Wie hängt die Stoffwechselrate bei der Ringeltaube von der Umgebungstemperatur ab?

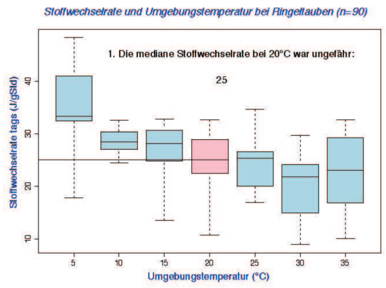
Daten
aus dem
AK Stoffwechselphysiologie

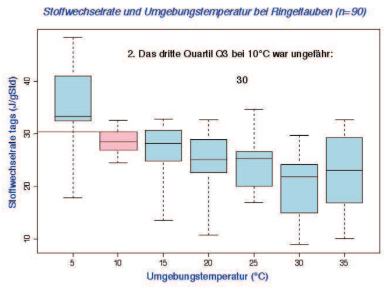
Prof. Prinzinger
Universität Frankfurt

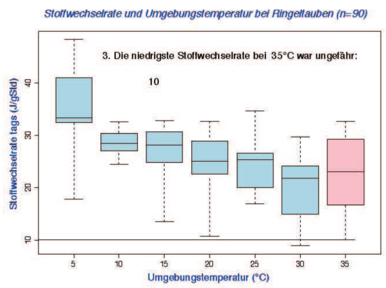


Klar:
Stoffwechselrate
höher
bei
tiefen Temperaturen

Vermutung:
Bei hohen Temperaturen
nimmt die Stoffwechselrate
wieder zu
(Hitzestress).

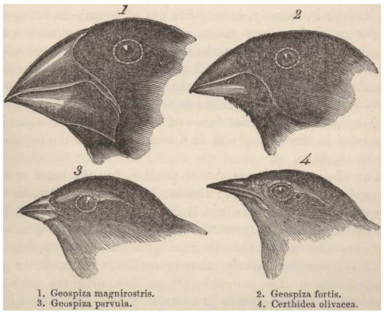






Charles Robert Darwin (1809-1882)

Darwin-Finken



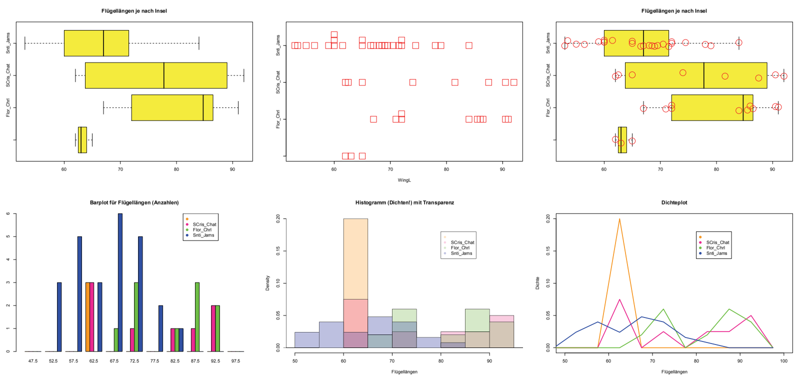
http:
//darwin-online.org.uk/graphics/Zoology_Illustrations.html

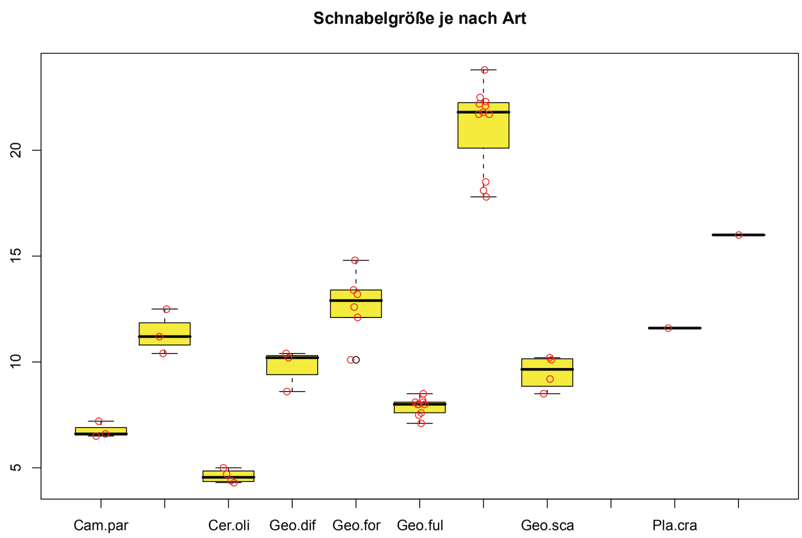
Darwins Finken-Sammlung

Sulloway, F.J. (1982) The Beagle collections of Darwin's Finches (Geospizinae). *Bulletin of the British Museum (Natural History), Zoology series* **43**: 49-94.

- http://datadryad.org/repo/handle/10255/dryad.154

Flügellängen der Darwin-Finken





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 verwenden
5. Vorsicht mit Tricks wie 3D oder halbtransparenten Farben
6. Jeder Datensatz ist anders; keine Patentrezepte

Es ist oft möglich,
das Wesentliche
an einer Stichprobe

mit ein paar Zahlen
zusammenzufassen.

Wesentlich:

1. Wie groß?

Lageparameter

2. Wie variabel?

Streuungsparameter

Eine Möglichkeit
kennen wir schon
aus dem Boxplot:

Lageparameter

Der Median

Streuungsparameter

Der Quartilabstand ($Q_3 - Q_1$)

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

Der Median ist
das 50%-Quantil
der Daten.

Die Quartile

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

Q_1 ist das
25%-Quantil
der Daten.

Die Quartile

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

Q_3 ist das
75%-Quantil
der Daten.

Am häufigsten werden benutzt:

Lageparameter

Der Mittelwert $\overline{x}$

Streuungsparameter

Die Standardabweichung s

Der Mittelwert

(engl. *mean*)

NOTATION:

Wenn die Beobachtungen

$x_1, x_2, x_3, \ldots, x_n$

heißen,

schreibt man oft

$\overline{x}$

für den Mittelwert.

DEFINITION:

Mittelwert = $\dfrac{\text{Summe der Messwerte}}{\text{Anzahl der Messwerte}}$ $\dfrac{\text{Summe}}{\text{Anzahl}}$

Der Mittelwert von $x_1, x_2, \ldots, x_n$ als Formel:

$$\begin{aligned}\overline{x} &= (x_1 + x_2 + \cdots + x_n)/n \\ &= \frac{1}{n}\sum_{i=1}^{n} x_i\end{aligned}$$

Beispiel:

$$x_1 = 3,\ x_2 = 0,\ x_3 = 2,\ x_4 = 3,\ x_5 = 1$$

$$\overline{x} = \text{Summe}/\text{Anzahl}$$

$$\overline{x} = (3 + 0 + 2 + 3 + 1)/5$$

$$\overline{x} = 9/5$$

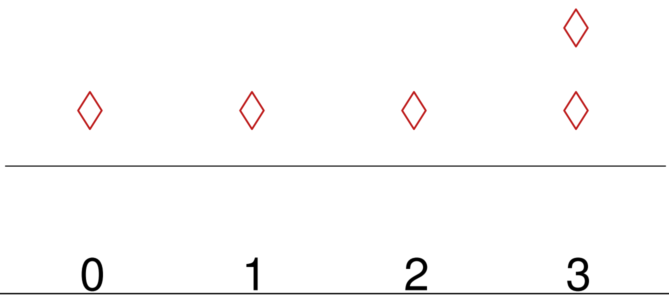
$$\overline{x} = 1{,}8$$

Geometrische Bedeutung des Mittelwerts:

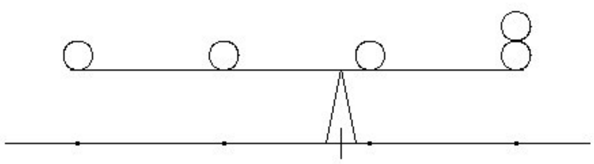
Der Schwerpunkt

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

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



$m = 1{,}8$?

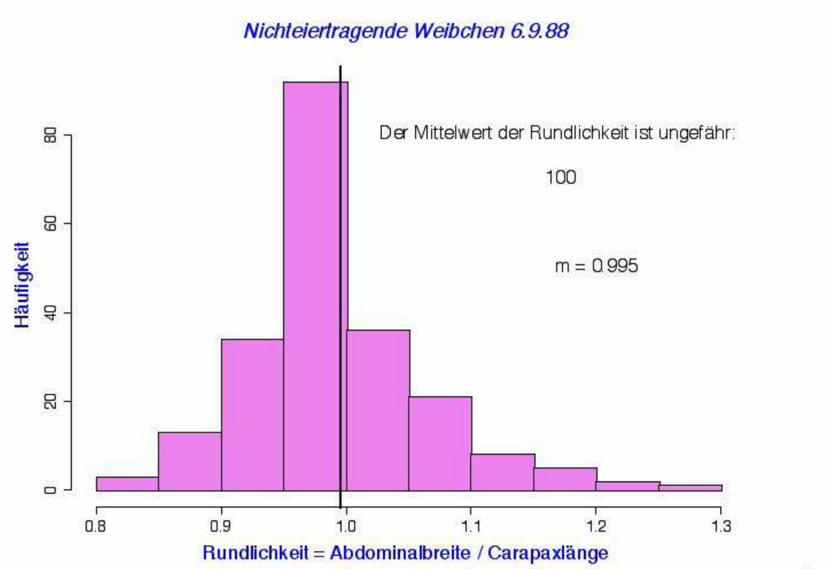


richtig

Beispiel: *Galathea intermedia*

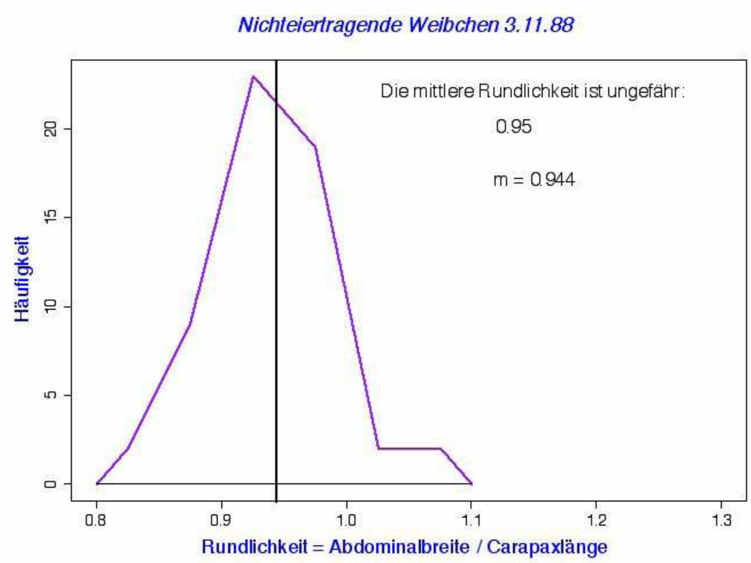
„Rundlichkeit“
:=
Abdominalbreite / Carapaxlänge

Vermutung:
Rundlichkeit nimmt
bei Geschlechtsreife zu



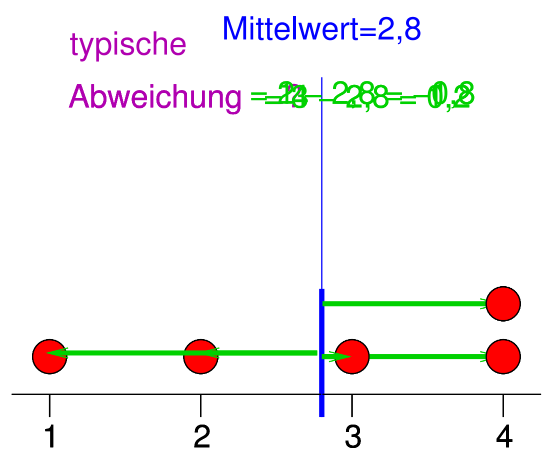
Beispiel:

3.11.88



Die Standardabweichung

Wie weit weicht
eine typische Beobachtung
vom
Mittelwert
ab ?



Die Standardabweichung σ (“sigma”)
[auch *SD* von engl. *standard deviation*]
ist ein

etwas komisches

gewichtetes Mittel
der Abweichungsbeträge

und zwar

$$\sigma = \sqrt{\text{Summe}(\text{Abweichungen}^2)/n}$$

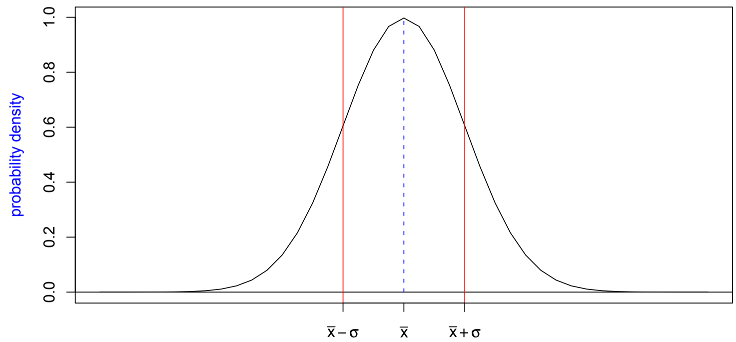
Die Standardabweichung von $x_1, x_2, \ldots, x_n$ als Formel:

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

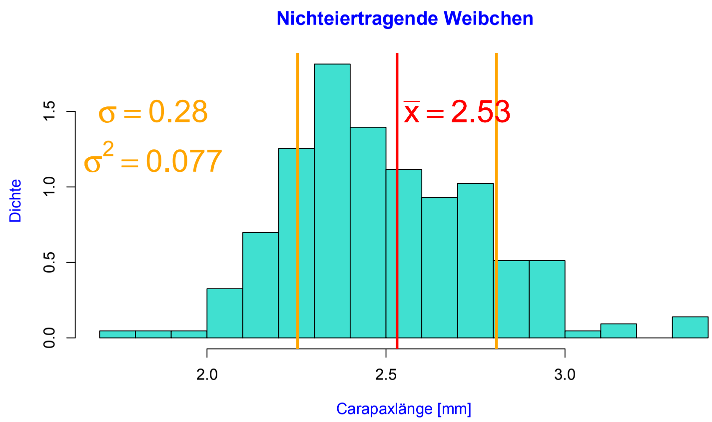
$\sigma^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^2$ heißt Varianz.

Faustregel für die Standardabweichung

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



Standardabweichung der Carapaxlängen nichteiertragender Weibchen vom 6.9.88



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

Varianz der Carapaxlängen nichteiertragender Weibchen vom 6.9.88

Alle Carapaxlängen im Meer: $\mathcal{X} = (X_1, X_2, \dots, X_N)$.
Carapaxlängen in unserer Stichprobe: $\mathcal{S} = (S_1, S_2, \dots, S_{n=215})$
Stichprobenvarianz:

$$\sigma_{\mathcal{S}}^2 = \frac{1}{n}\sum_{i=1}^{215}(S_i - \overline{S})^2 \approx 0{,}0768$$

Können wir 0,0768 als Schätzwert für die Varianz σ_X^2 in der ganzen Population verwenden?
Ja, können wir machen. Allerdings ist $\sigma_{\mathcal{S}}^2$ im Durchschnitt um den Faktor $\frac{n-1}{n}$ $(= 214/215 \approx 0,995)$ kleiner als $\sigma_{\mathcal{X}}^2$

Varianzbegriffe

Varianz in der Population: $\sigma_X^2 = \frac{1}{N}\sum_{i=1}^{N}(X_i - \overline{X})^2$

Stichprobenvarianz: $\sigma_{\mathcal{S}}^2 = \frac{1}{n}\sum_{i=1}^{n}(S_i - \overline{S})^2$
korrigierte Stichprobenvarianz:

$$\begin{aligned} s^2 &= \frac{n}{n-1}\sigma_{\mathcal{S}}^2 \\ &= \frac{n}{n-1}\cdot\frac{1}{n}\cdot\sum_{i=1}^{n}(S_i - \overline{S})^2 \\ &= \frac{1}{n-1}\cdot\sum_{i=1}^{n}(S_i - \overline{S})^2 \end{aligned}$$

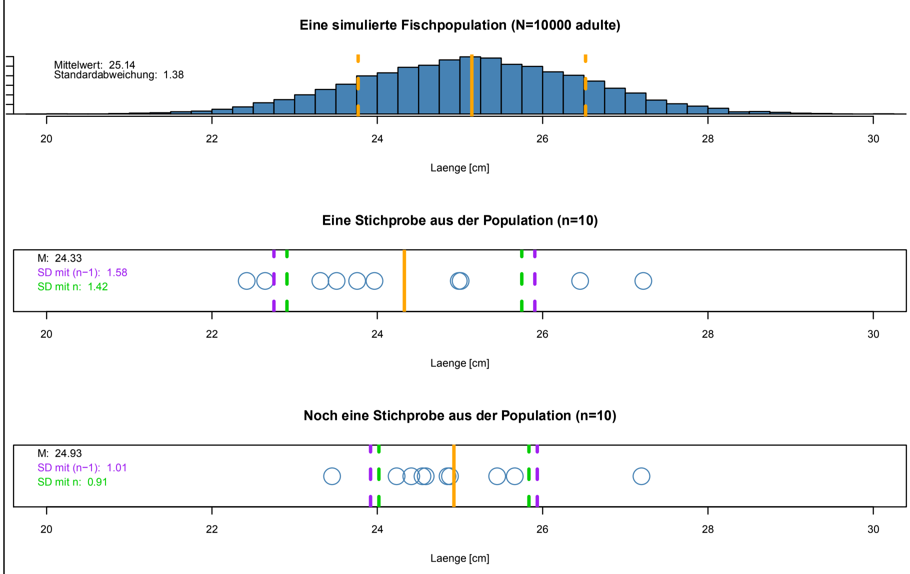
Mit “Standardabweichung von $\mathcal{S}$” ist meistens das korrigierte s gemeint.

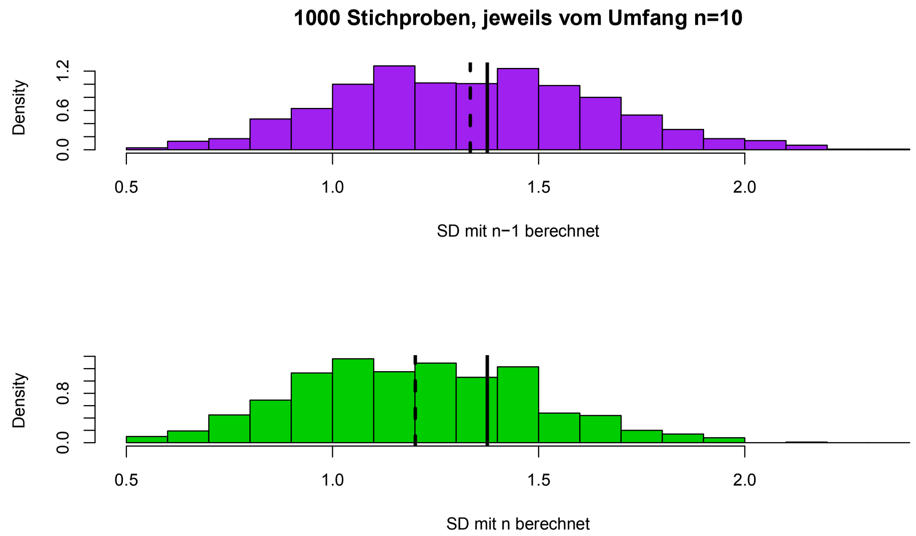
$\overline{x} = 10/5 = 2$

						Summe
x	1	3	0	5	1	10
$x - \overline{x}$	-1	1	-2	3	-1	0
$(x - \overline{x})^2$	1	1	4	9	1	16

$$s^2 = \text{Summe}((x - \overline{x})^2)/(n-1) = 16/(5-1) = 4$$

$$s = 2$$





σ mit n oder $n-1$ berechnen?

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

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

Wenn es sich bei $x_1, \ldots, x_n$ um eine Stichprobe handelt (wie meistens in der Statistik), sollten Sie die Formel

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

verwenden.

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. Glauben Sie uns also nicht alle Datenpunkte.

Bachstelzen fressen Dungfliegen

Räuber

Bachstelze (White Wagtail)
Motacilla alba alba

image (c) by Artur Mikołajewski

Beute

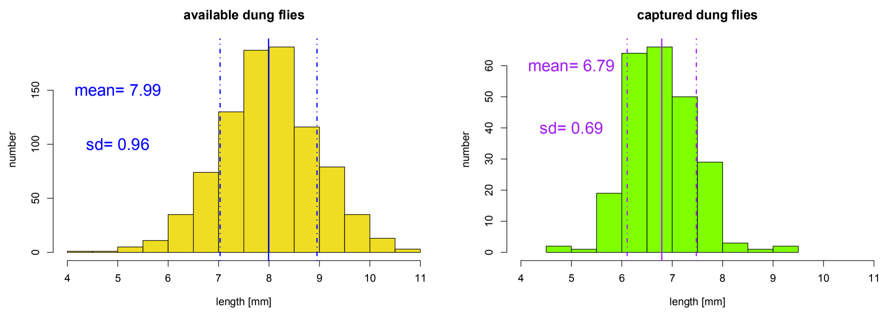
Gelbe Dungfliege
Scatophaga stercoraria

image (c) by Viatour Luc

Vermutung

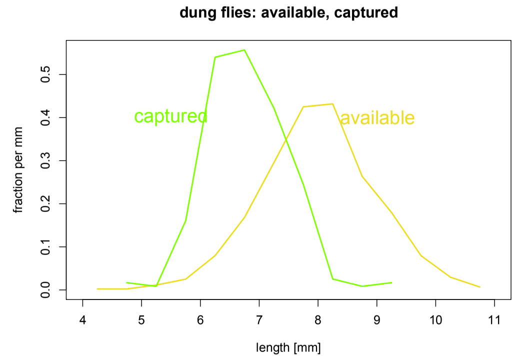
- Die Fliegen sind unterschiedlich groß
- Effizienz für die Bachstelze = Energiegewinn / Zeit zum Fangen und fressen
- Laborexperimente lassen vermuten, dass die Effizienz bei 7mm großen Fliegen maximal ist.

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



Vergleich der Größenverteilungen

	captured		available
Mittelwert	6.29	$<$	7.99
Standardabweichung	0.69	$<$	0.96



Interpretation

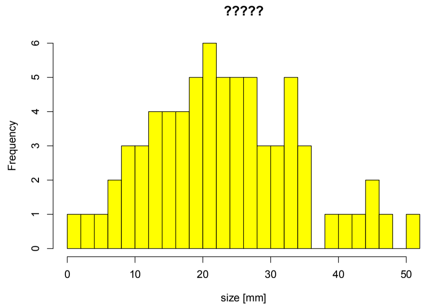
Die Bachstelzen bevorzugen Dungfliegen, die etwa 7mm groß sind.

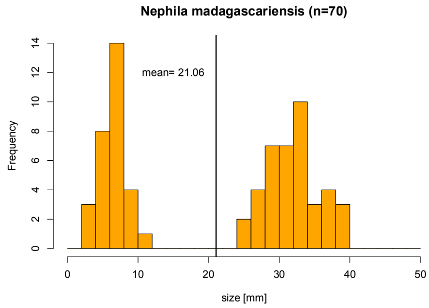
Hier waren die Verteilungen glockenförmig und es genügten 4 Werte (die beiden Mittelwerte und die beiden Standardabweichungen), um die Daten adäquat zu beschreiben.

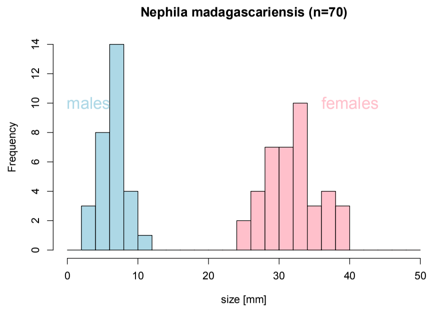
Nephila madagascariensis

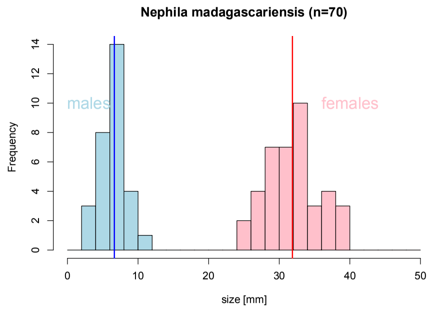
image (c) by Bernard Gagnon

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

Kupfertolerantes Rotes Straußgras

Rotes Straußgras
Agrostis tenuis
image (c) Kristian Peters

Kupfer
Cuprum
Hendrick met de Bles

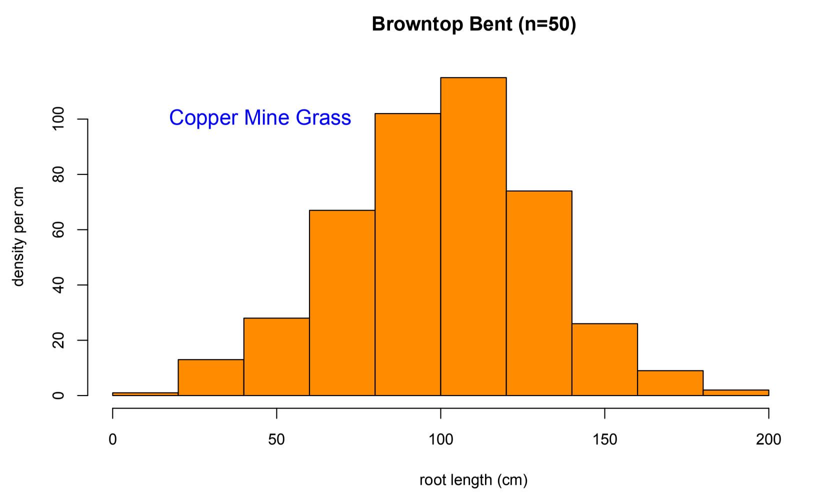
A.D. Bradshaw.
Population Differentiation in *agrostis tenius Sibth*. III. populations in varied environments.
New Phytologist, 59(1):92 – 103, 1960.

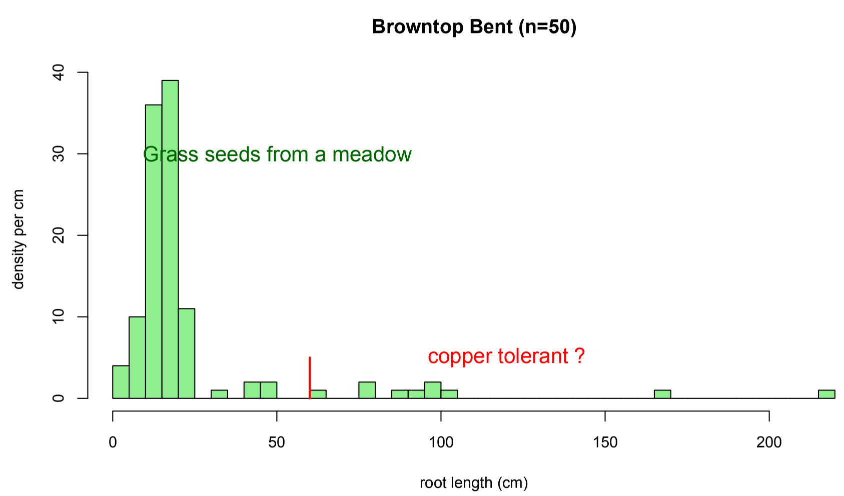
T. McNeilly and A.D Bradshaw.
Evolutionary Processes in Populations of Copper Tolerant Agrostis tenuis Sibth.
Evolution, 22:108–118, 1968.

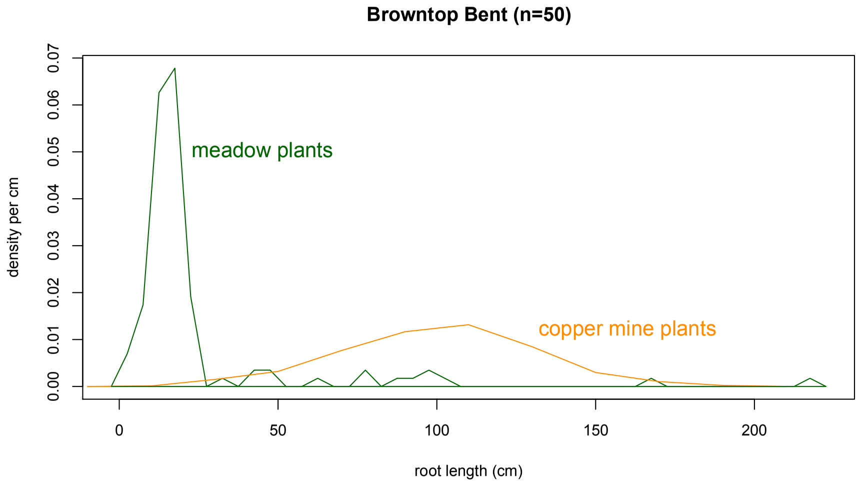
Wir verwenden hier wieder simulierte Daten, da die Originaldaten nicht zur Verfügung stehen.

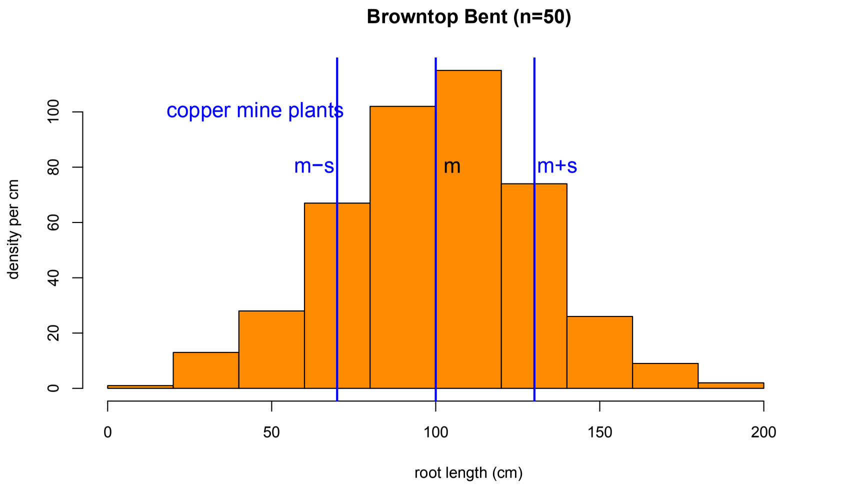
Anpassung an Kupfer?

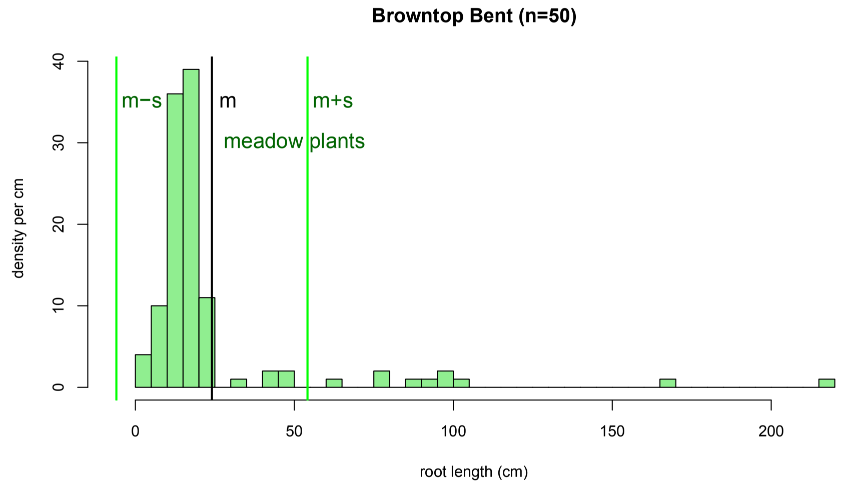
- Pflanzen, denen das Kupfer schadet, haben kürzere Wurzeln.
- Die Wurzellängen von Pflanzen aus der Umgebung von Kupferminen wird gemessen.
- Samen von unbelasteten Wiesen werden bei Kupferminen eingesäht.
- Die Wurzellängen dieser “Wiesenpflanzen” werden gemessen.


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
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)

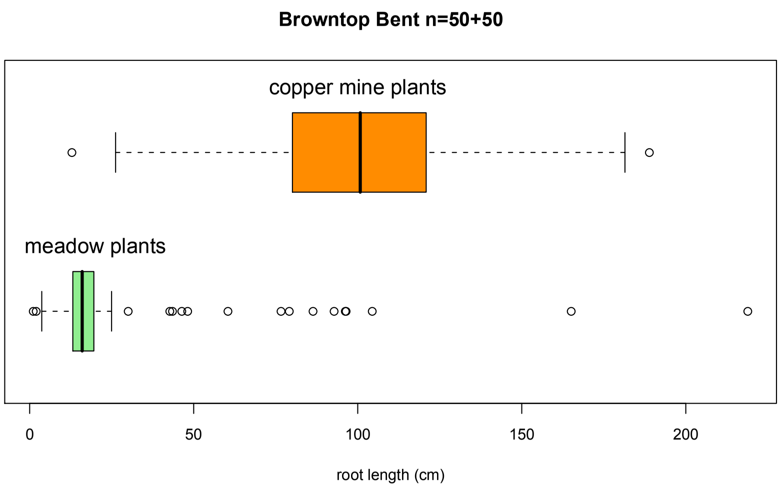


2/3 der Wurzellängen innerhalb [m-sd,m+sd]???? Nein!

Fazit des Straußgras-Beispiels

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

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



Schlussfolgerung

In der Biologie sind viele Datenverteilungen annähernd glockenförmig und können durch den Mittelwert und die Standardabweichung hinreichend beschrieben werden.

Es gibt aber auch Ausnahmen. Also:
Immer die Daten erst mal graphisch untersuchen!

Verlassen sie sich niemals allein auf numerische Kenngrößen!