Wahrscheinlichkeitsrechnung und Statistik für Biologen
Einführung: Deskriptive Statistik

Martin Hutzenthaler & Dirk Metzler

17. April 2012

Inhaltsverzeichnis

1 Einführung

1.1 Konzept und Quellen

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

Andrejs Dunkels

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.

Also: Statistik ist Datenanalyse mit Hilfe stochastischer Modelle.

Quellen

Wir danken Matthias Birkner für die intensive Zusammenarbeit beim Erstellen der ersten Version dieser Vorlesung sowie Brooks Ferebee, Gaby Schneider und Anton Wakolbinger für die Bereitstellung vieler Beispiele und Lehrmaterialien.

http://joguinf.informatik.uni-mainz.de/~birkner/
http://www.math.uni-frankfurt.de/~wakolbin/statbio/
http://ismi.math.uni-frankfurt.de/schneider/statbio0708.html

1.2 Plan

Plan der Vorlesung

Klassische Statistik

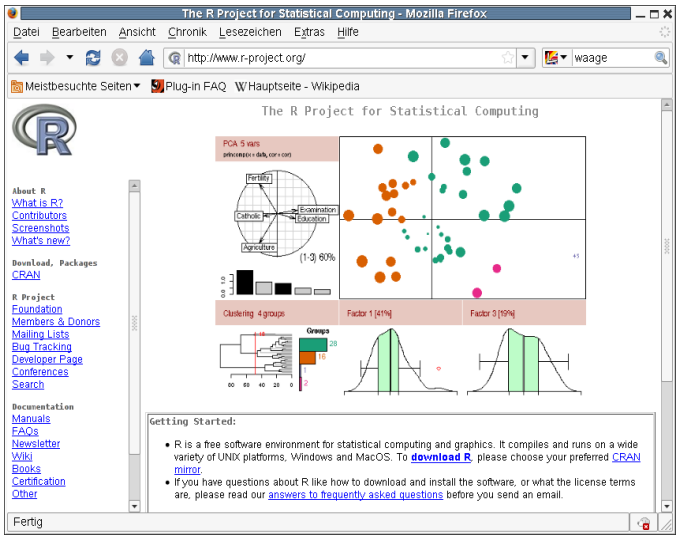
1. Beschreibende Statistik
2. Der Standardfehler
3. Der t-Test für gepaarte Stichproben
4. Der t-Test für unabhängige Stichproben
5. Häufigkeiten
6. Der Chi-Quadrat Test
7. Lineare Regression
8. Korrelation
9. Varianzanalyse (ANOVA)

Plan der Vorlesung

Weitere Themen

- Nichtparametrische Tests
- Diskriminanzanalyse
- Grundbegriffe der Wahrscheinlichkeitstheorie
- Parameterschätzung
- Moderne Anwendung: Analyse von Genexpressionsdaten (vielleicht)
- R

Statistik-Software R



http://www.r-project.org

Folien, R-Befehle, Quellen und Übungen

http://evol.bio.lmu.de/statgen/StatBiol/12SS

2 Ziele der deskriptiven (d.h. beschreibenden) Statistik

Beschreibende Statistik

Beschreibende Statistik: Ein erster Blick auf die Daten

3 Graphische Darstellungen

Beispiel

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

Crustaceensektion

Leitung: Dr. Michael Türkay

Charybdis acutidens TÜRKAY 1985

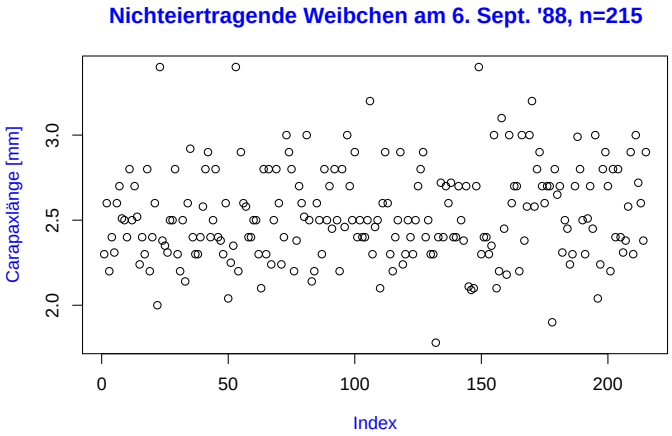
Der Springkrebs

Galathea intermedia

Helgoländer Tiefe Rinne, Fang vom 6.9.1988

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

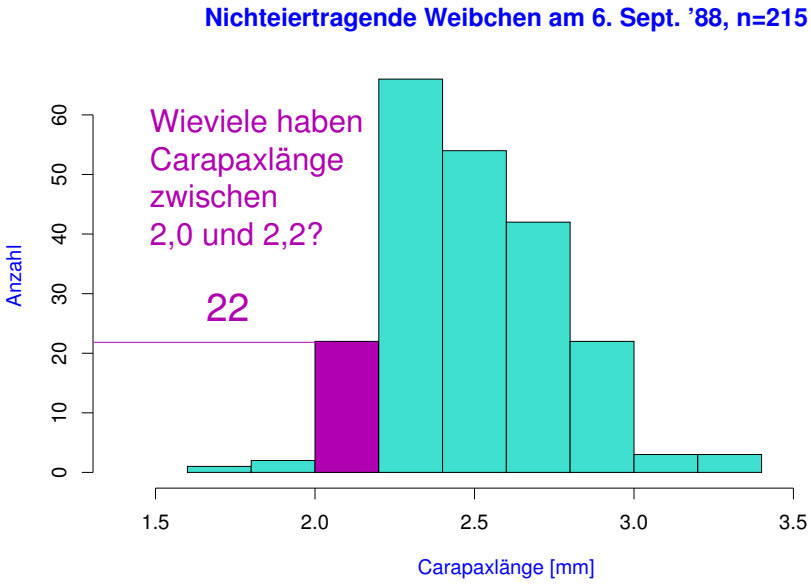
2,9	3,0	2,9	2,5	2,7	2,9	2,9	3,0
3,0	2,9	3,4	2,8	2,9	2,8	2,8	2,4
2,8	2,5	2,7	3,0	2,9	3,2	3,1	3,0
2,7	2,5	3,0	2,8	2,8	2,8	2,7	3,0
2,6	3,0	2,9	2,8	2,9	2,9	2,3	2,7
2,6	2,7	2,5	.	.	.	.	.



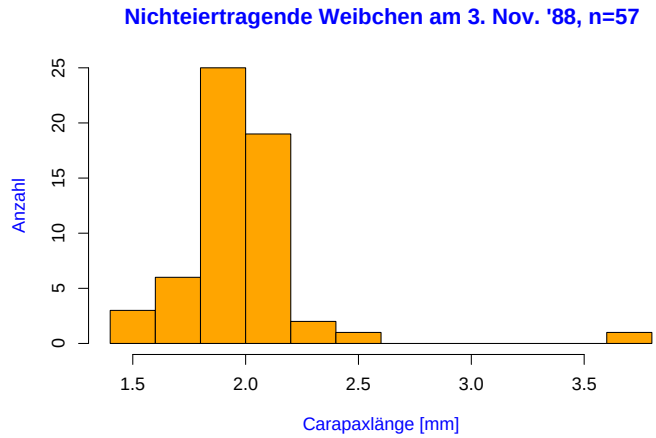
3.1 Histogramme und Dichtepolygone

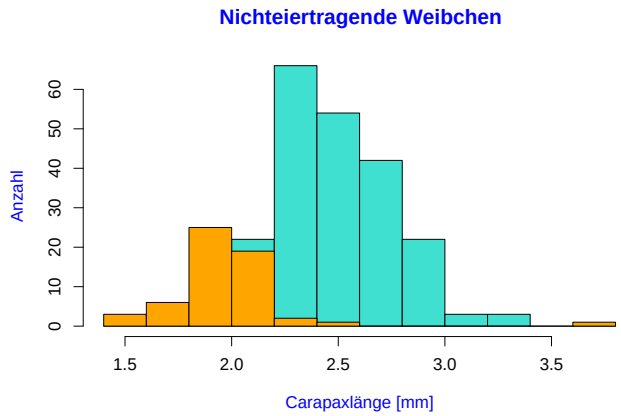
Eine Möglichkeit der graphischen Darstellung:

das Histogramm



Analoge Daten zwei Monate später (3.11.88):

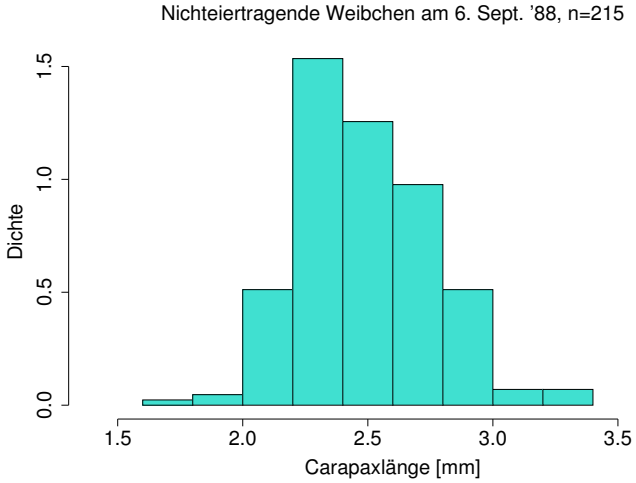




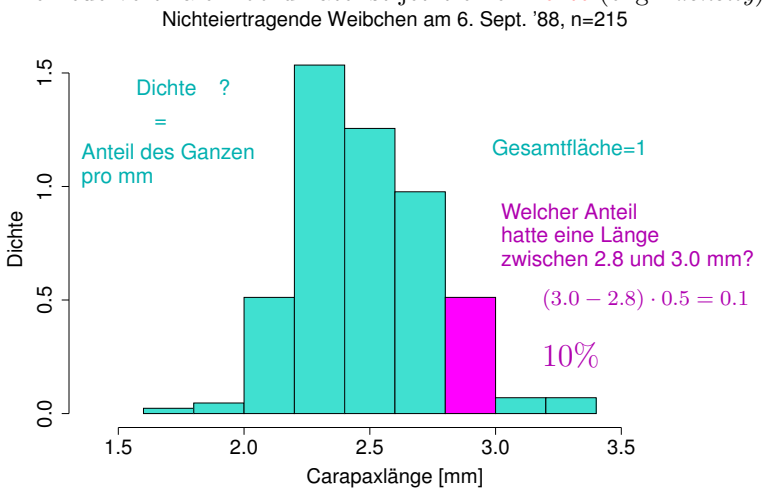
Vergleich der beiden Verteilungen

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

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

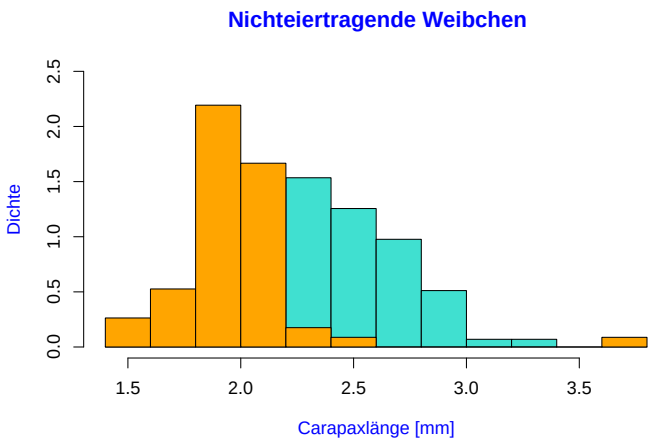


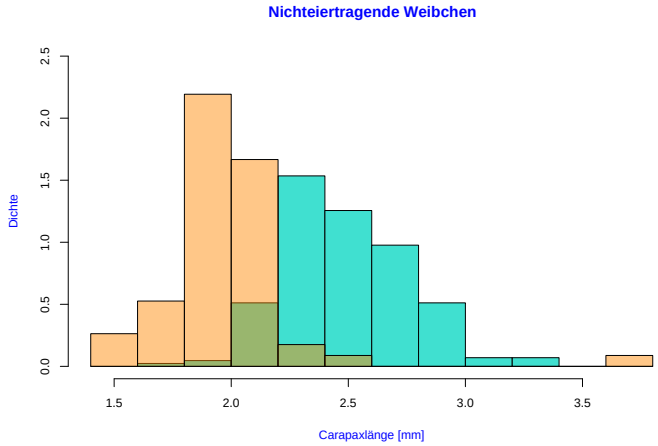
Die neue vertikale Koordinate ist jetzt eine Dichte (engl. *density*).

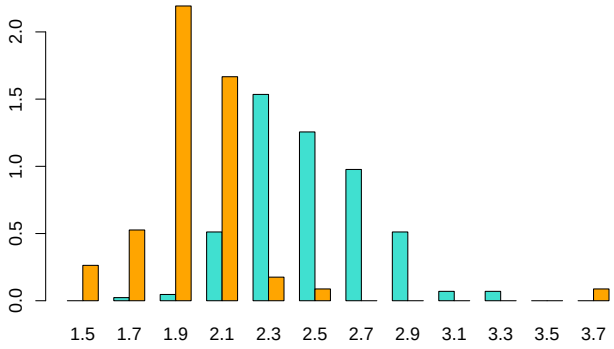


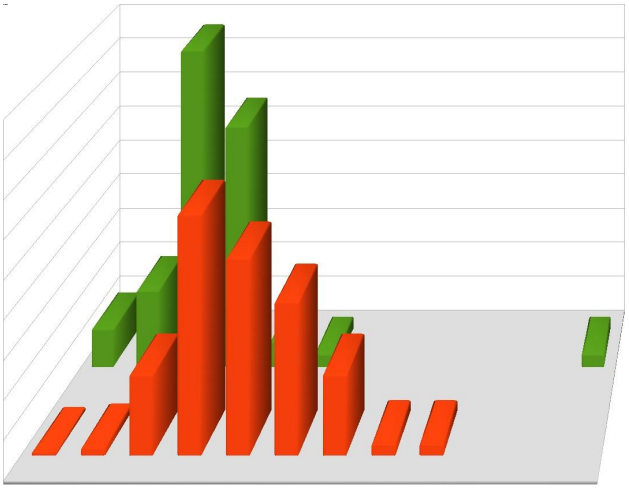
Die beiden Histogramme sind jetzt vergleichbar, denn sie haben dieselbe Gesamtfläche:

Versuche, die Histogramme zusammen zu zeigen:









Unser Rat an Sie:

Wenn Sie Schauwerbegestalter(in) sind:

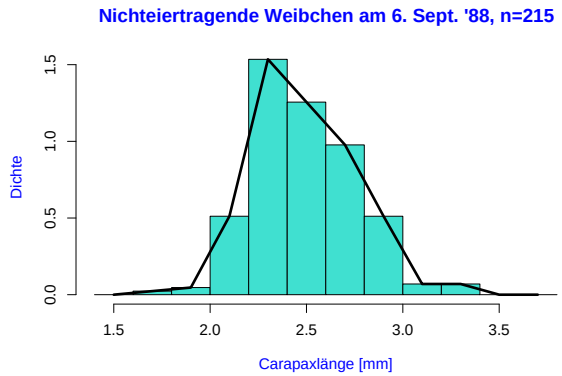
Beeindrucken Sie Jung und Alt mit total abgefahrenen 3D-Plots!

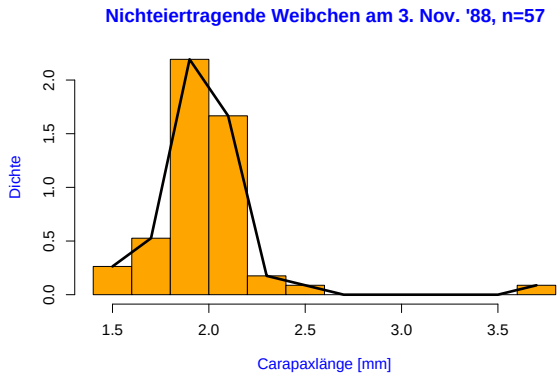
Wenn Sie Wissenschaftler(in) werden wollen:

Bevorzugen Sie einfache und klare 2D-Darstellungen.

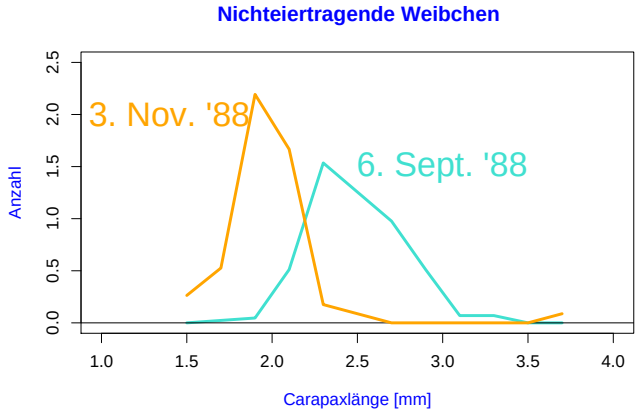
Problem: Histogramme kann man nicht ohne weiteres in demselben Graphen darstellen, weil sie einander überdecken würden.

Einfache und klare Lösung: Dichtepolygone

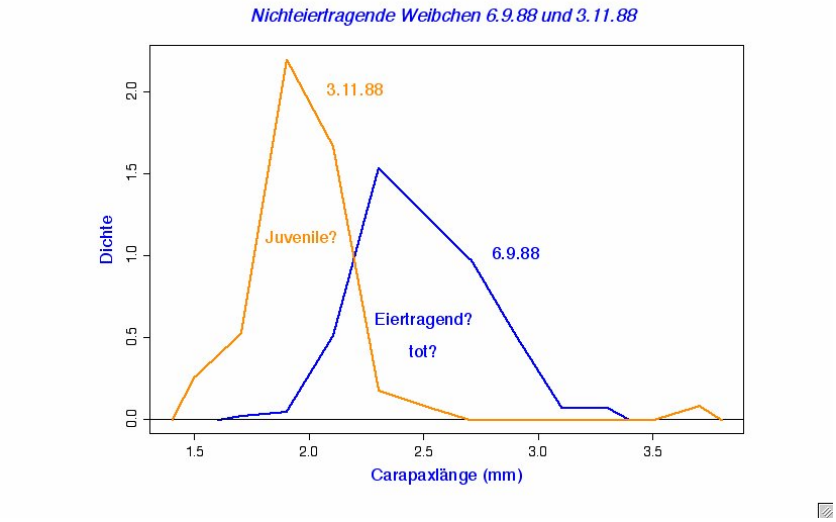




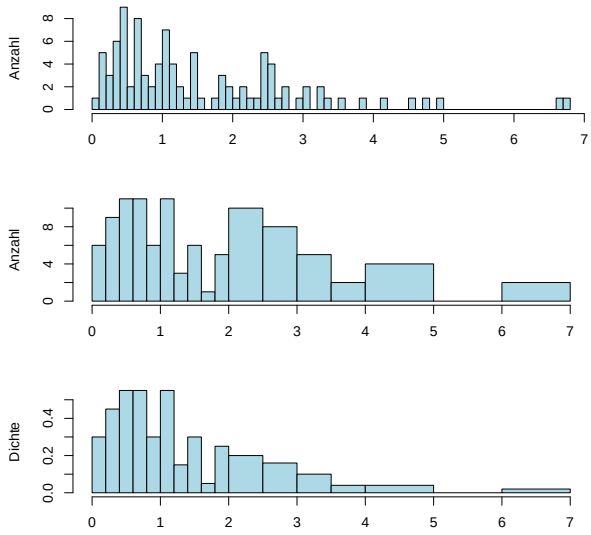
Zwei und mehr Dichtepolygone in einem Plot



Biologische Interpretation der Verschiebung?

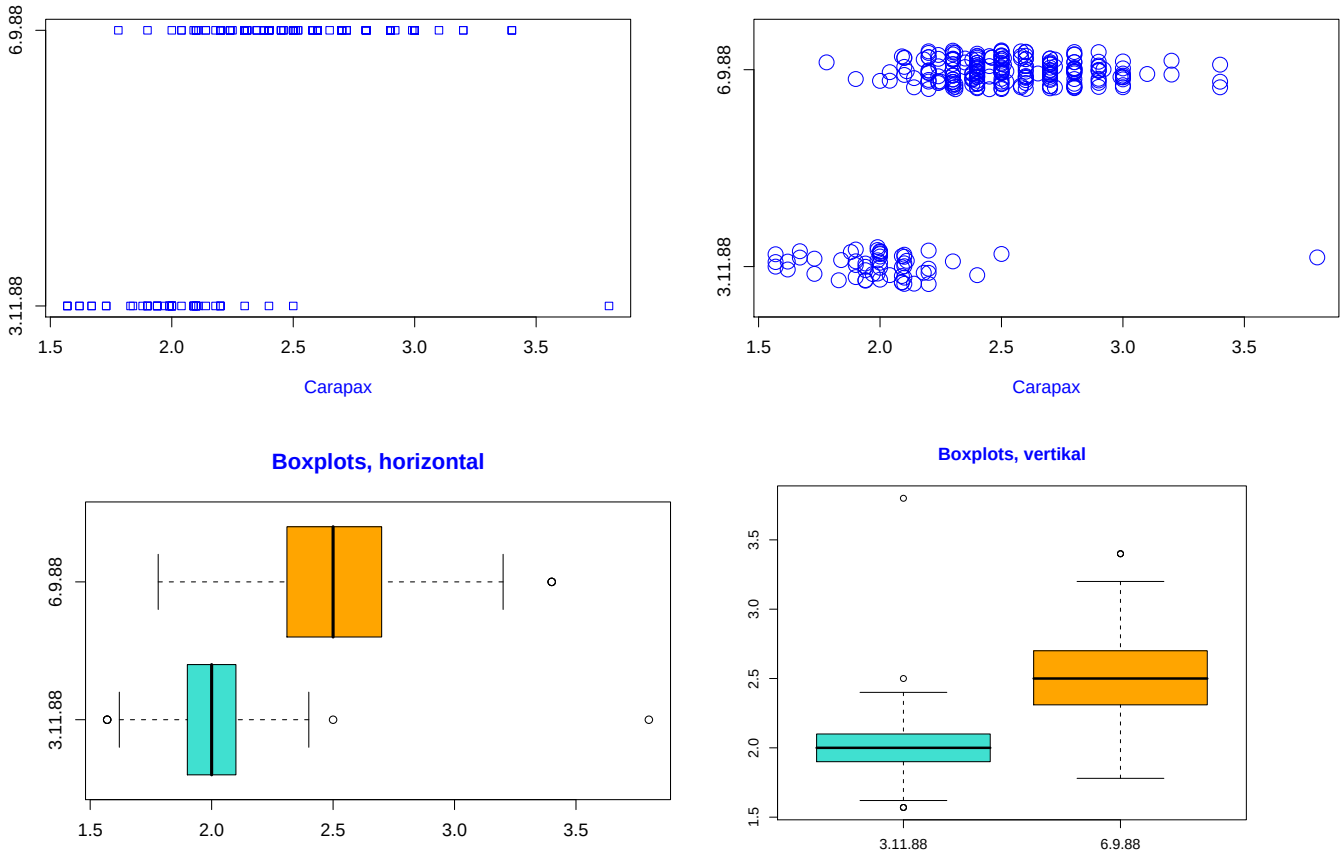


Anzahl vs. Dichte



Also: Bei Histogrammen mit ungleichmäßiger Unterteilung immer Dichten verwenden!

3.2 Stripcharts



Histogramme und Dichtepolygone geben ein ausführliches Bild eines Datensatzes. Manchmal zu ausführlich.

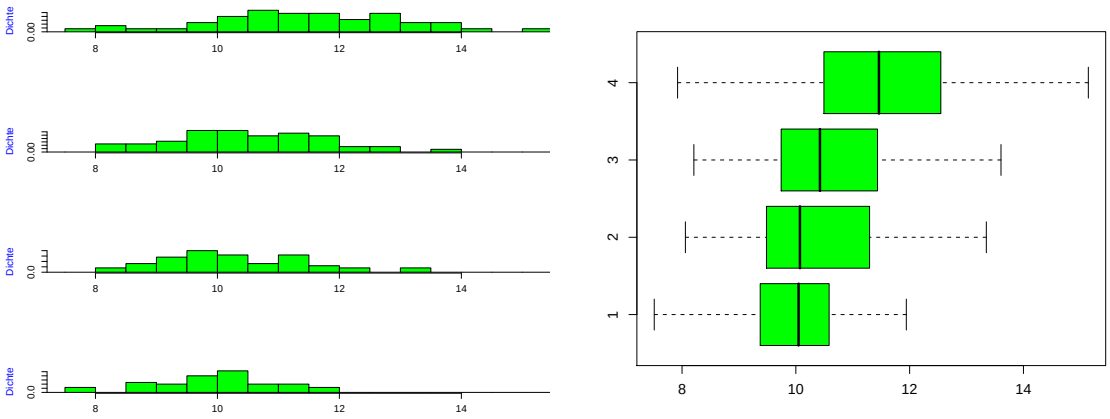
3.3 Boxplots

Zu viel Information erschwert den Überblick

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

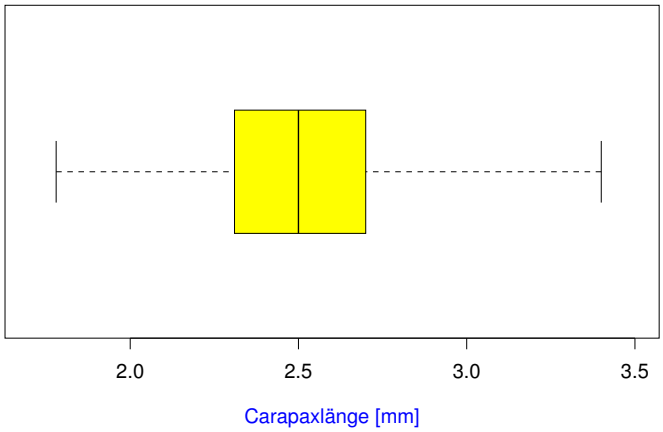
Wald?

Beispiel:
Vergleich von mehreren Gruppen



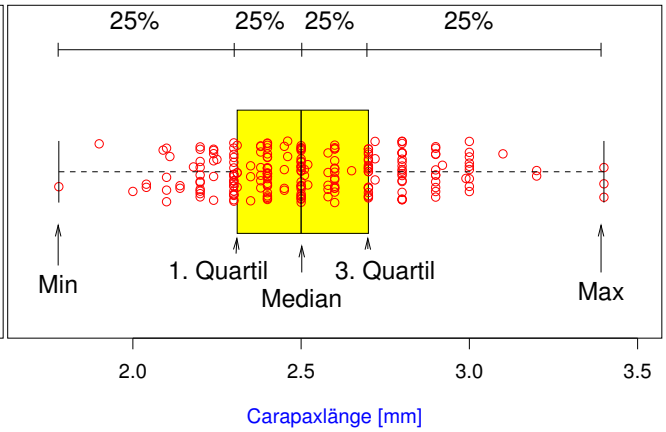
Der Boxplot

Boxplot, einfache Ausführung



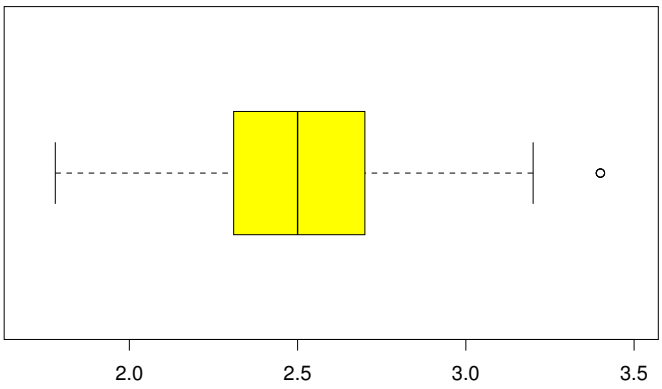
Carapaxlänge [mm]

Boxplot, einfache Ausführung



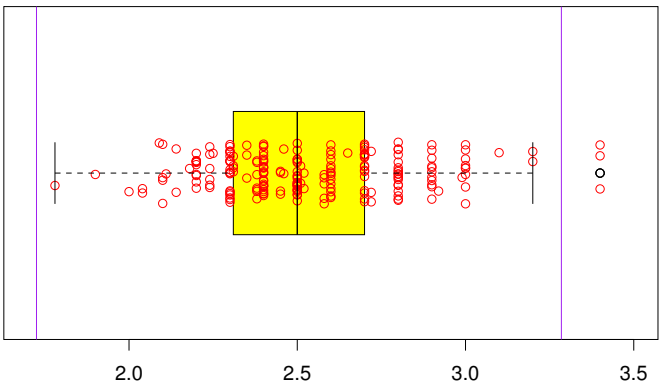
Carapaxlänge [mm]

Boxplot, Standardausführung

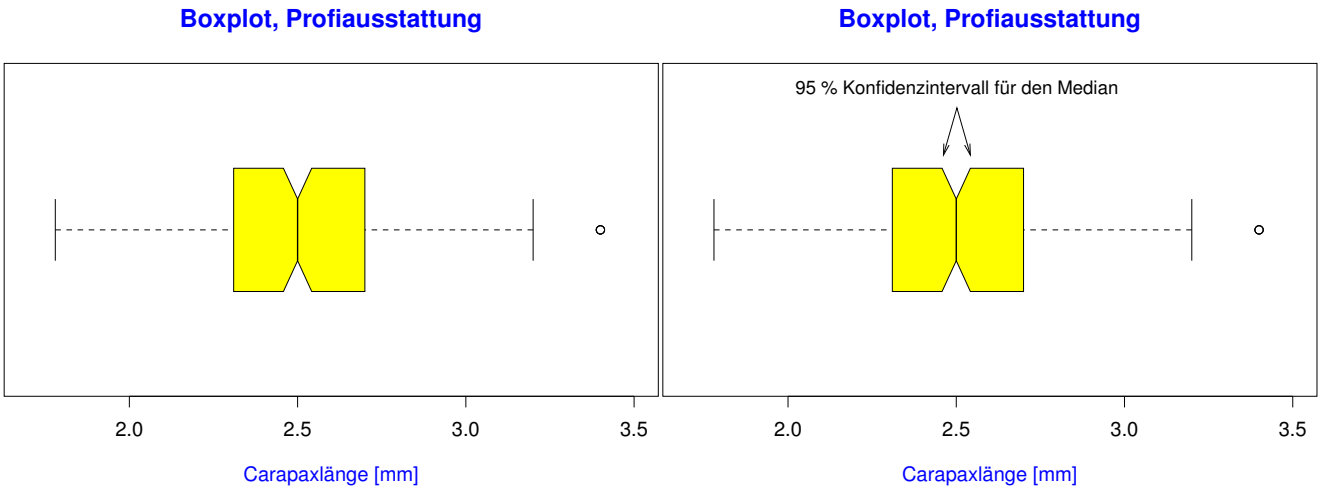


Carapaxlänge [mm]

Boxplot, Standardausführung



Carapaxlänge [mm]



3.4 Beispiel: Ringeltaube

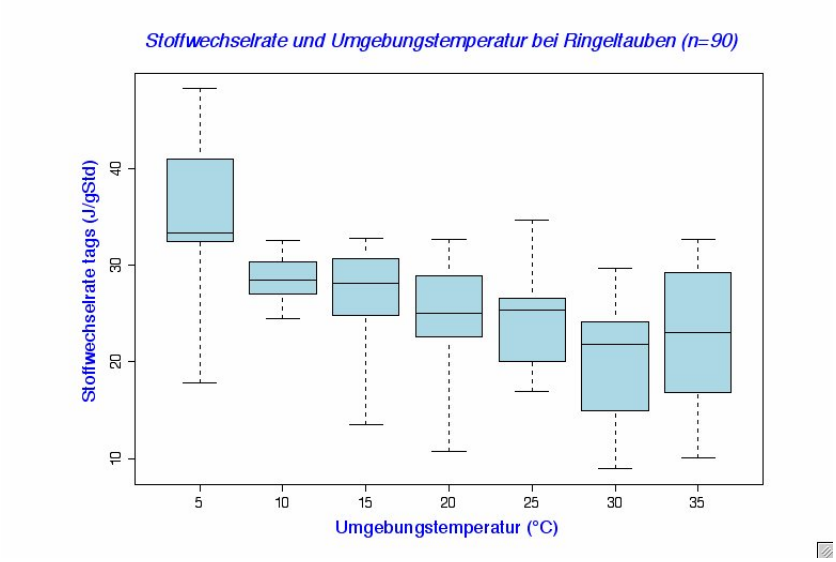
Beispiel:

Die Ringeltaube

Palumbus palumbus

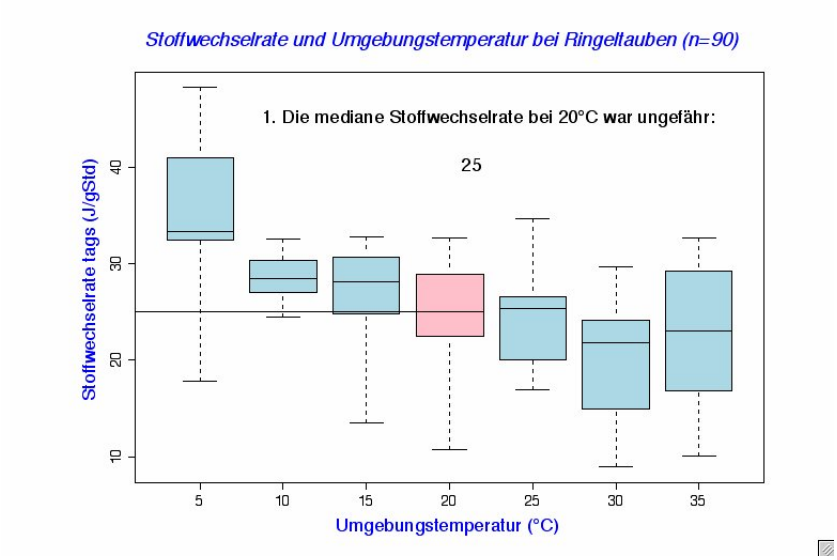
Wie hängt die Stoffwechselrate bei der Ringeltaube von der Umgebungstemperatur ab?

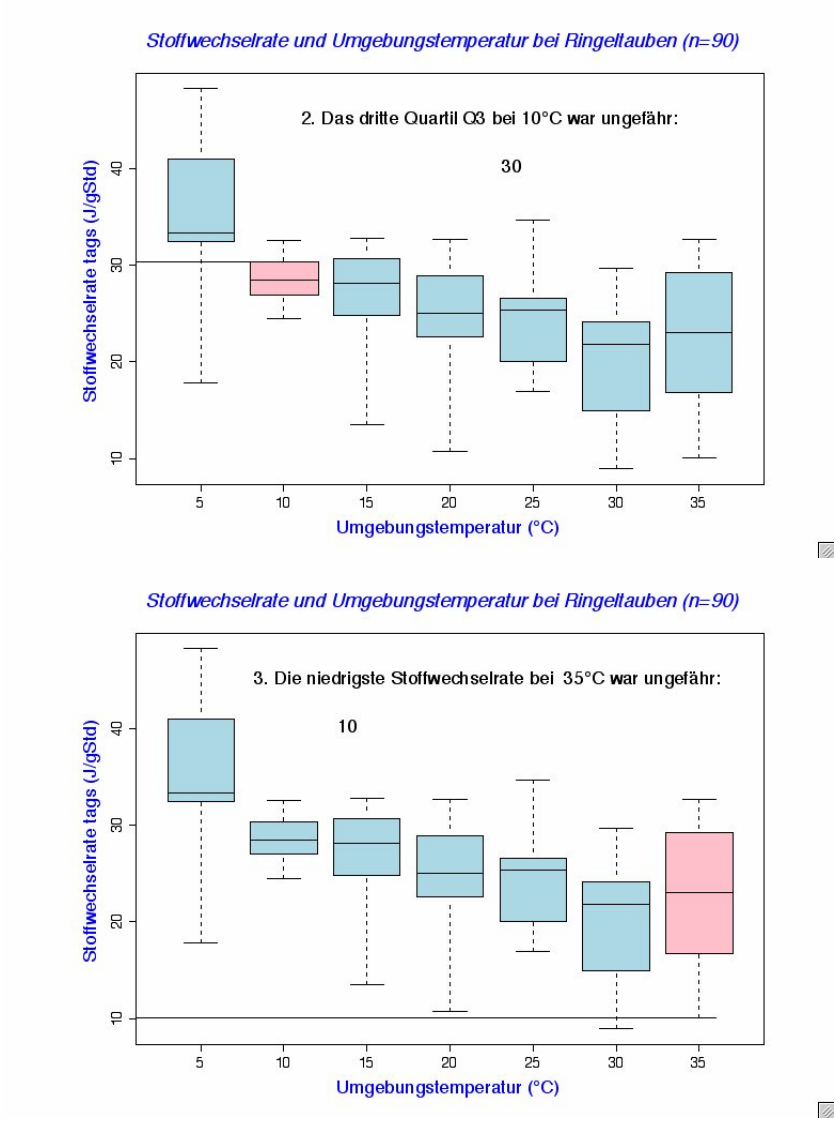
Daten aus dem AK Stoffwechselphysiologie
Prof. Prinzinger Universität Frankfurt



Klar: Stoffwechselrate *höher* bei *tiefen* Temperaturen

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

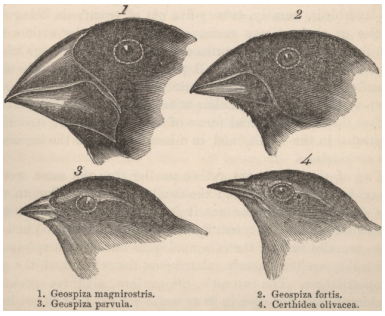




3.5 Beispiel: Darwin-Finken

Charles Robert Darwin (1809-1882)

Darwin-Finken



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

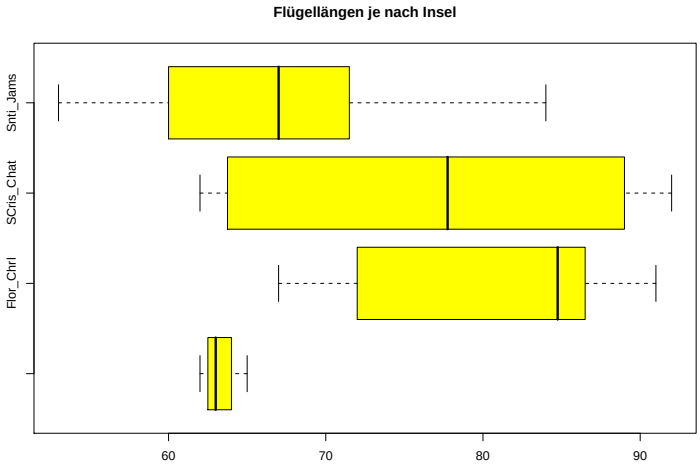
Darwins Finken-Sammlung

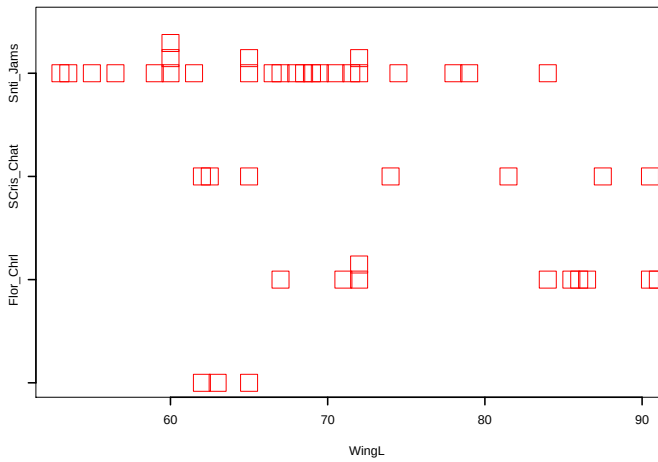
Literatur

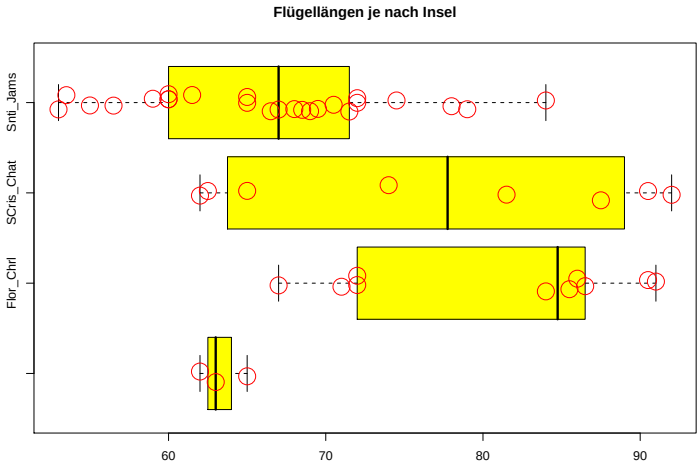
[1] Sulloway, F.J. (1982) The Beagle collections of Darwin's Finches (Geospizinae). *Bulletin of the British Museum (Natural History), Zoology series* **43**: 49-94.

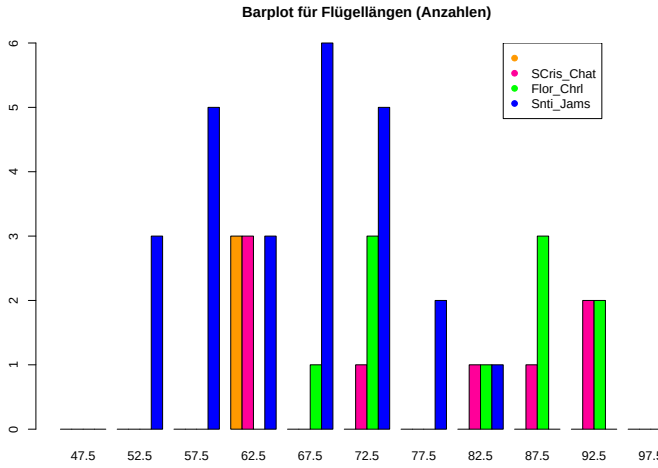
[2] http://datadryad.org/repo/handle/10255/dryad.154

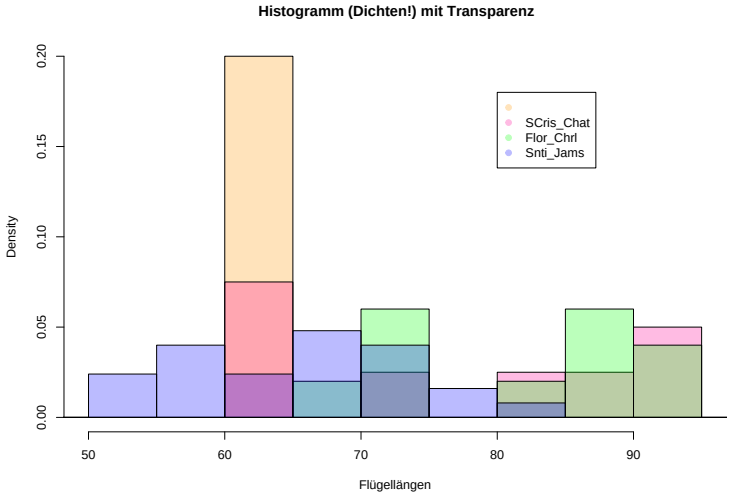
Flügellängen der Darwin-Finken

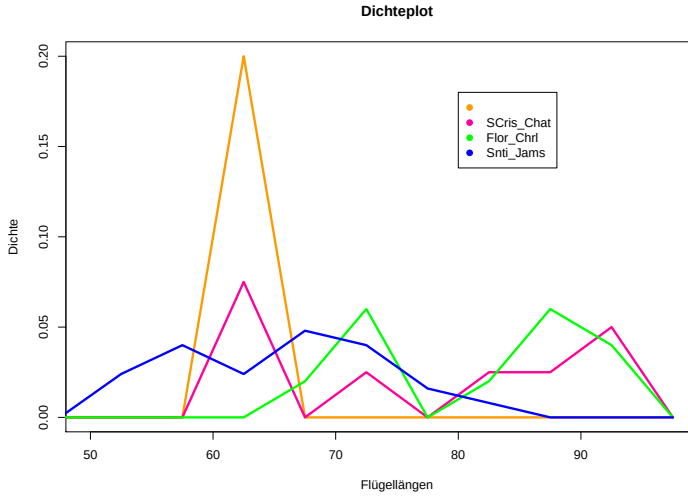

Flügellängen je nach Insel
Snti_Jams
SCris_Chat
Flor_Chrl
60
70
80
90

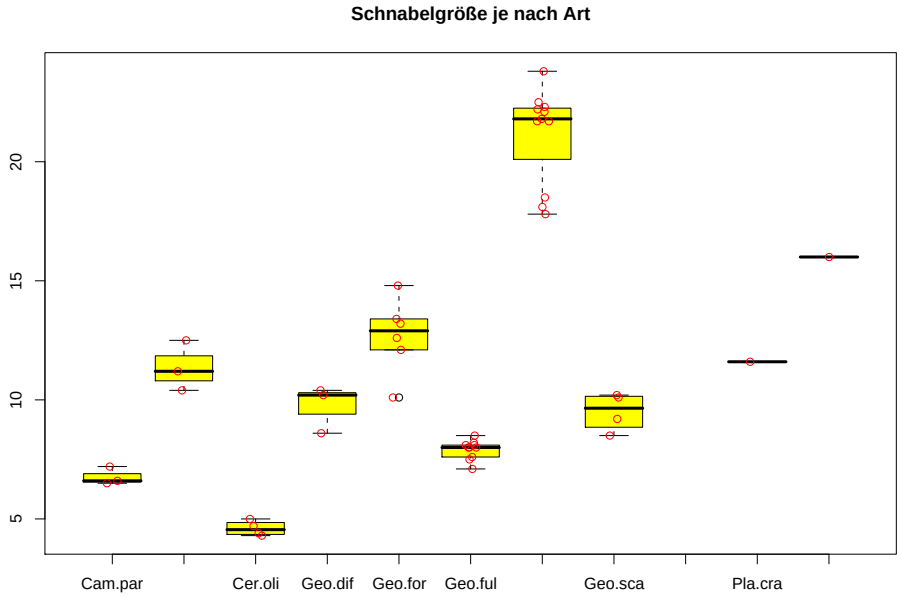

Snti_Jams
SCris_Chat
Flor_Chrl
60
70
80
90
WingL


Flügellängen je nach Insel
Snti_Jams
SCris_Chat
Flor_Chrl
60
70
80
90


Barplot für Flügellängen (Anzahlen)
SCris_Chat
Flor_Chrl
Snti_Jams
0
1
2
3
4
5
6
47.5
52.5
57.5
62.5
67.5
72.5
77.5
82.5
87.5
92.5


Histogramm (Dichten!) mit Transparenz
Density
0.00
0.05
0.10
0.15
0.20
SCris_Chat
Flor_Chrl
Snti_Jams
50
60
70
80
90
Flügellängen


Dichteplot
Dichte
0.00
0.05
0.10
0.15
0.20
SCris_Chat
Flor_Chrl
Snti_Jams
50
60
70
80
90
Flügellängen



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

4 Statistische Kenngrößen

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 $(Q_3 - Q_1)$

4.1 Median und andere Quartile

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.

4.2 Mittelwert und Standardabweichung

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:

$$\text{Mittelwert[1ex]} = \text{[1ex]} \frac{\text{Summe der Messwerte}}{\text{Anzahl der Messwerte}}$$

$$\frac{\text{Summe}}{\text{Anzahl}}$$

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:

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

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

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

$\overline{x} = 9/5$

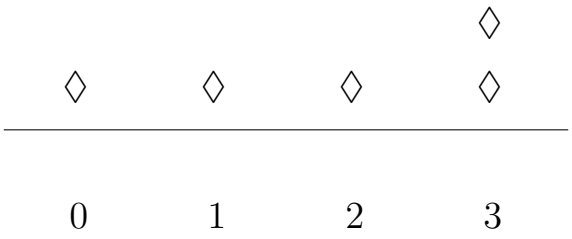
$\overline{x} = 1{,}8$

Geometrische Bedeutung des Mittelwerts:

Der Schwerpunkt

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

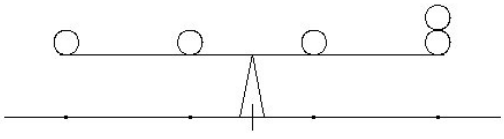
Wo muß der Drehpunkt sein, damit die Waage im Gleichgewicht ist?

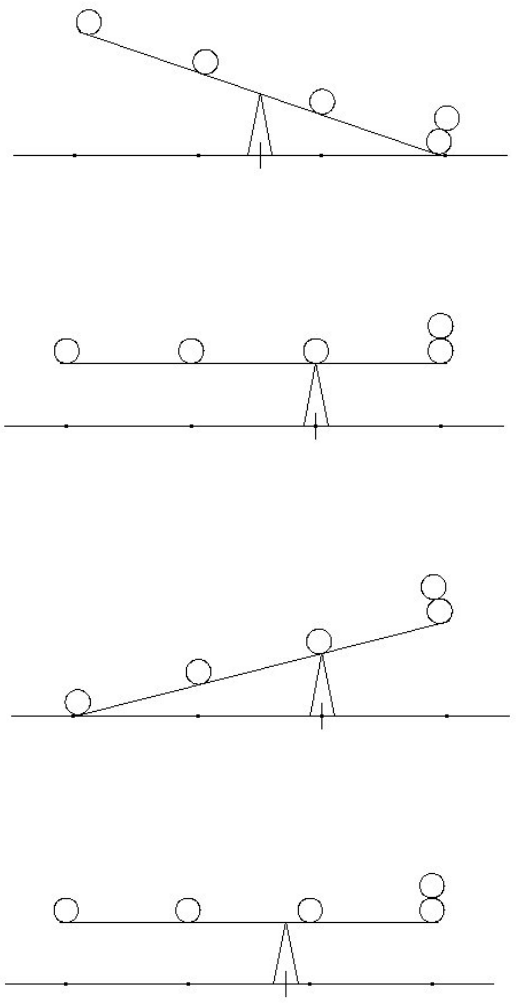


$m = 1{,}5$?

$m = 2$?

$m = 1{,}8$?





zu klein

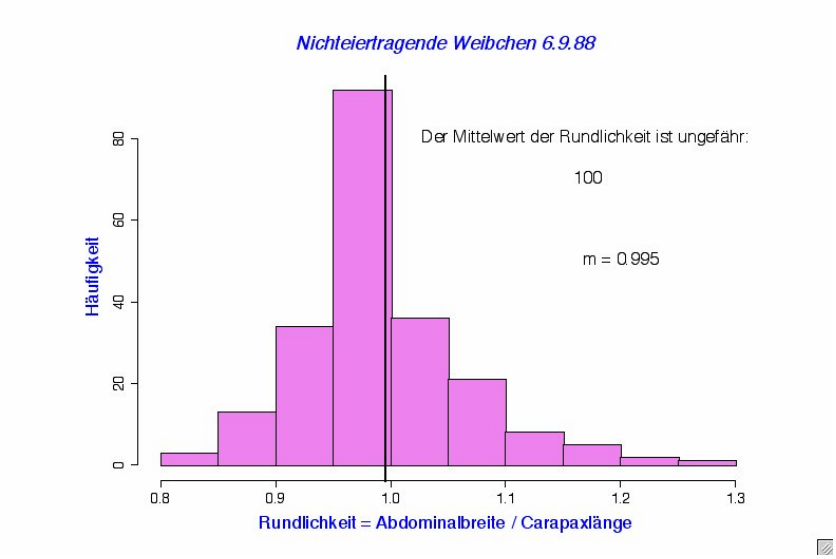
zu groß

richtig

Beispiel: *Galathea intermedia*

„Rundlichkeit“ := Abdominalbreite / Carapaxlänge

Vermutung: Rundlichkeit nimmt bei Geschlechtsreife zu

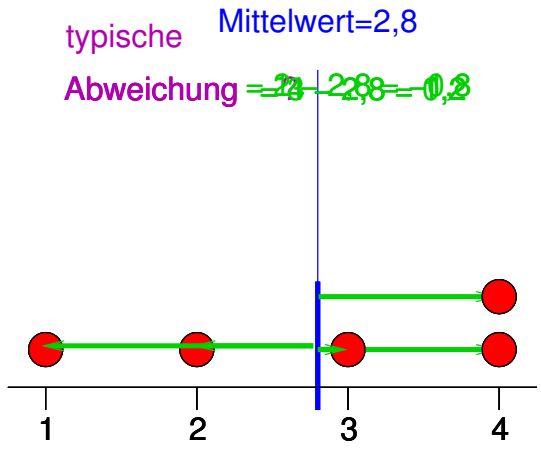


Beispiel:

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

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

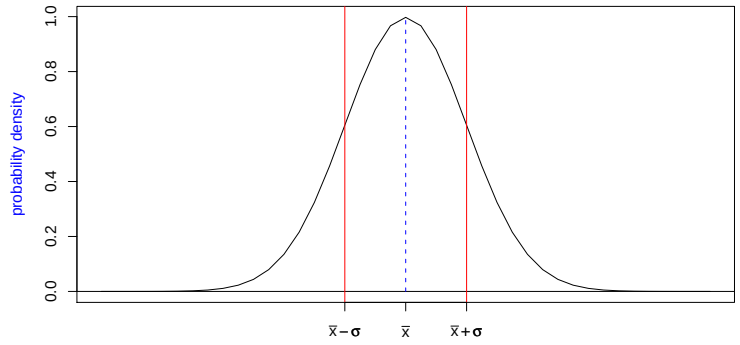
Die *Standardabweichung* von $x_1, x_2, \ldots, x_n$ als Formel:

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

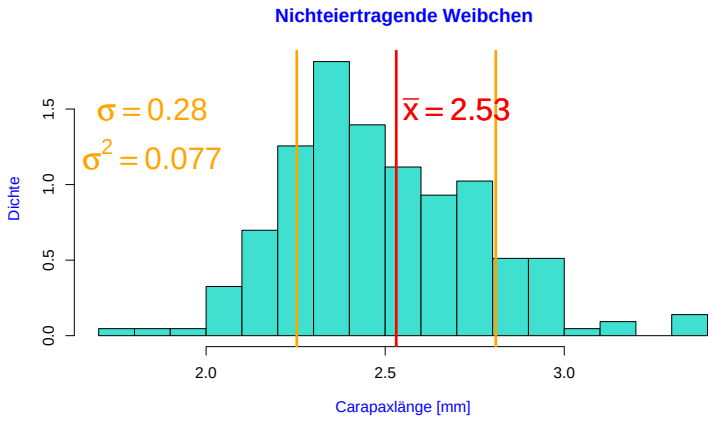
$\sigma^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^2$ heißt *Varianz*.

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

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

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

Varianzbegriffe

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

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

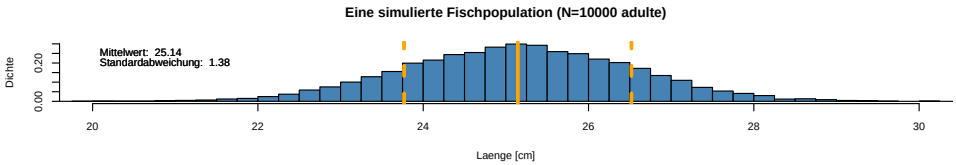
$$\begin{aligned} s^2 &= \frac{n}{n-1}\sigma^2_{\mathcal{S}} \\ &= \frac{n}{n-1} \cdot \frac{1}{n} \cdot \sum_{i=1}^{n}(S_i - \overline{S})^2 \\ &= \frac{1}{n-1} \cdot \sum_{i=1}^{n}(S_i - \overline{S})^2 \end{aligned}$$

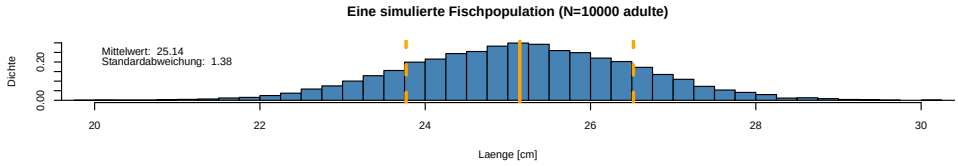
Mit "Standardabweichung von $\mathcal{S}$" ist meistens das korrigierte s gemeint.

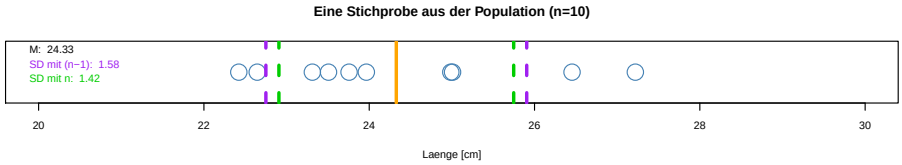
Beispiel Die Daten $\overline{x} =? \overline{x} = 10/5 = 2$

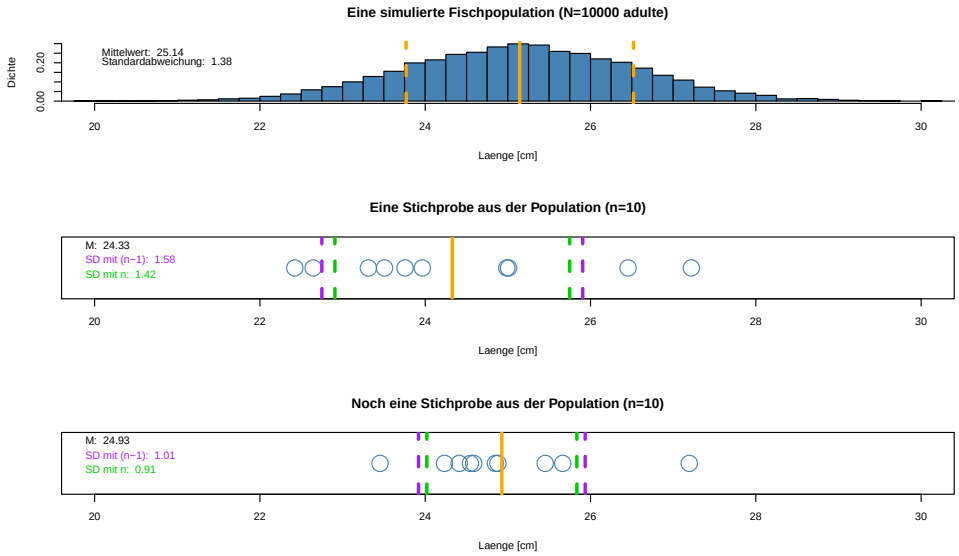
						Summe
x	1	3	0	5	1	10
$x - \overline{x}$	-1	1	-2	3	-1	0
$(x - \overline{x})^2$	1	1	4	9	1	16

$$\begin{aligned} s^2 &= \mathit{Summe}\big((x - \overline{x})^2\big)/(n-1) \\ &= 16/(5-1) = 4 \\ s &= 2 \end{aligned}$$

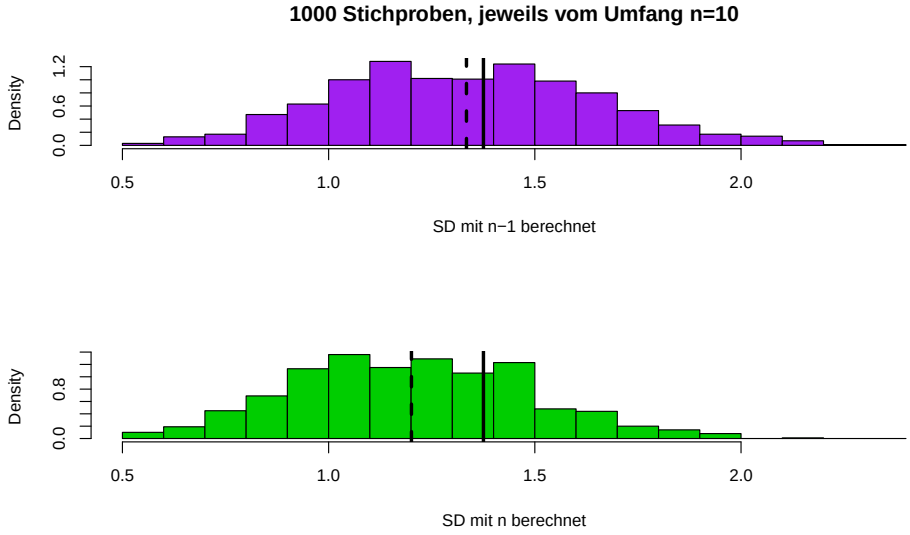

Eine simulierte Fischpopulation (N=10000 adulte)
Mittelwert: 25.14
Standardabweichung: 1.38
Dichte
0.00
0.20
20
22
24
26
28
30
Laenge [cm]


Eine simulierte Fischpopulation (N=10000 adulte)
Mittelwert: 25.14
Standardabweichung: 1.38
Dichte
0.00
0.20
20
22
24
26
28
30
Laenge [cm]


Eine Stichprobe aus der Population (n=10)
M: 24.33
SD mit (n−1): 1.58
SD mit n: 1.42
20
22
24
26
28
30
Laenge [cm]



Die folgenden Histogramme zeigen die Standardabweichungen, die aus 1000 verschiedenen Stichproben aus der selben Verteilung geschätzt wurden. Die durchgezogenen Linien stellen die tatsächliche Standardabweichung der Verteilung dar, die gestrichelten Linien die Mittelwerte der geschätzten Standardabweichugen.



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.

5 Vom Sinn und Unsinn von Mittelwerten

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.

Literatur

[BTH08] 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.

5.1 Beispiel: Wählerische Bachstelzen

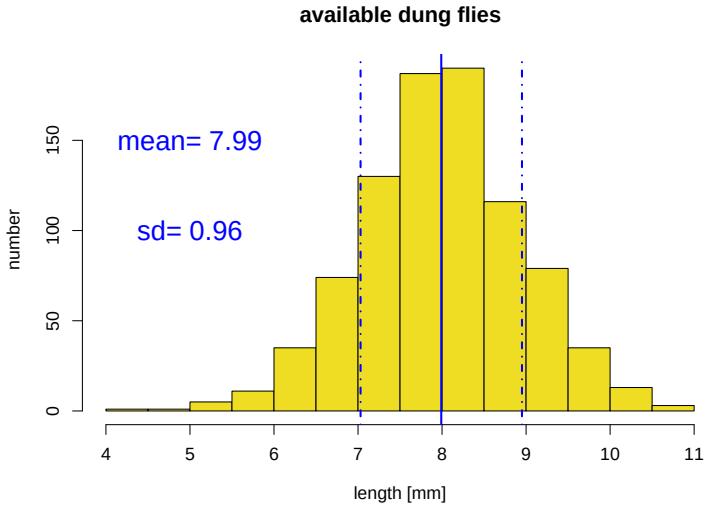
Bachstelzen fressen Dungfliegen

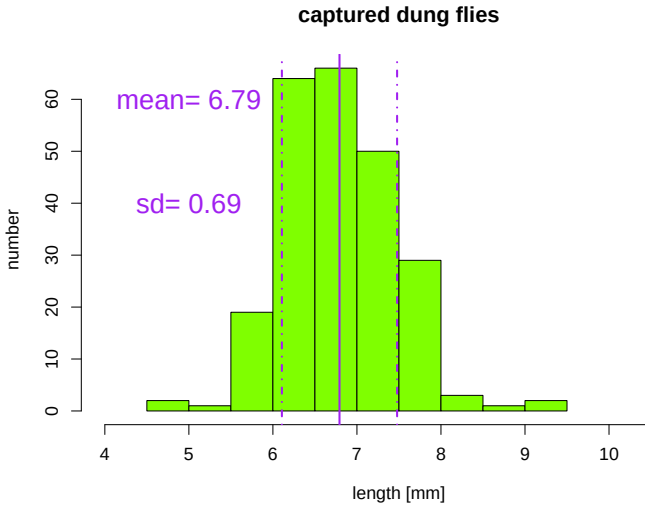
Vermutung

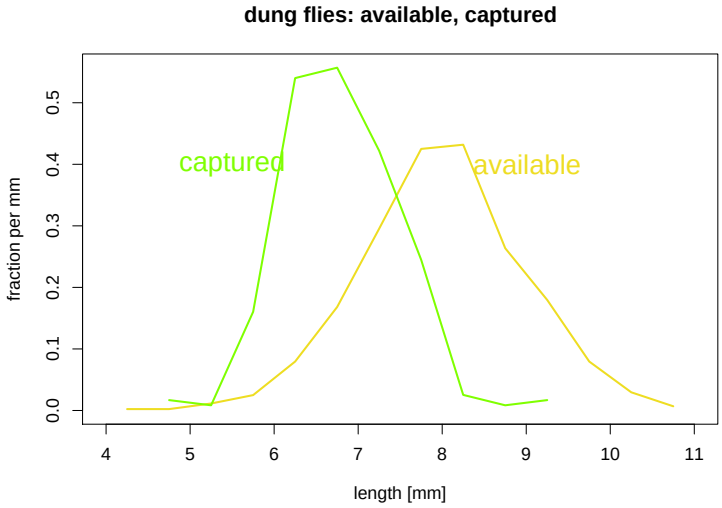
- Die Fliegen sind unterschiedlich groß
- Effizienz für die Bachstelze = Energiegewinn / Zeit zum Fangen und fressen
- Laborexperimente lassen vermuten, dass die Effizienz bei 7mm großen Fliegen maximal ist.

Literatur

[Dav77] 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

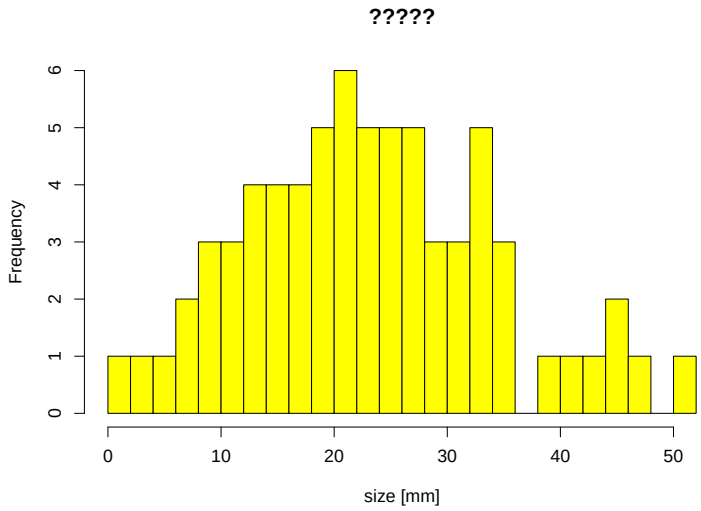
Die Bachstelzen bevorzugen Dungfliegen, die etwa 7mm groß sind.

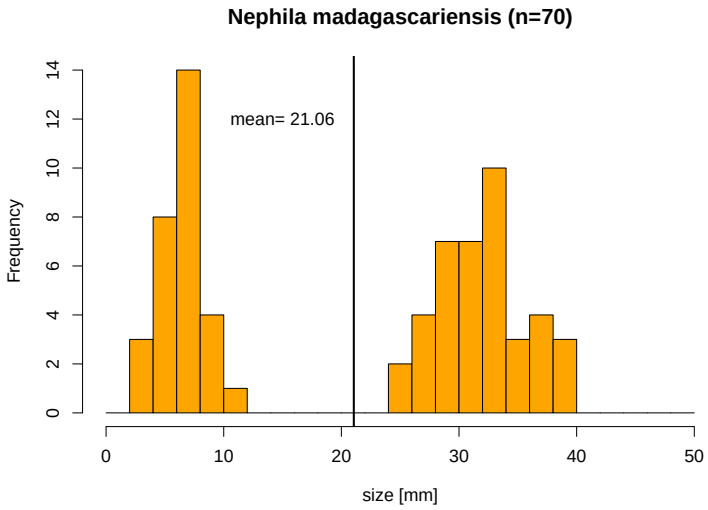
Hier waren die Verteilungen glockenförmig und es genügten 4 Werte (die beiden Mittelwerte und die beiden Standardabweichungen), um die Daten adäquat zu beschreiben.

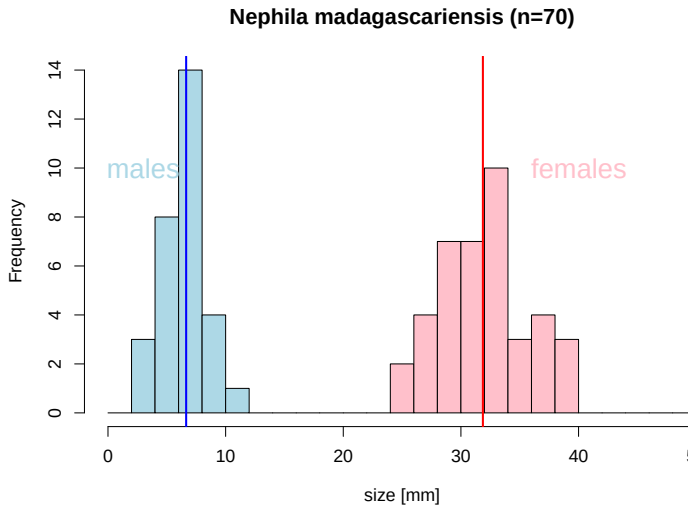
5.2 Beispiel: Spiderman & Spiderwoman

Simulated Data:
Eine Stichprobe von 70 Spinnen
Mittlere Größe: 21,06 mm

Standardabweichung der Größe: 12,94 mm







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.

5.3 Beispiel: Kupfertoleranz beim Roten Straußgras

Literatur

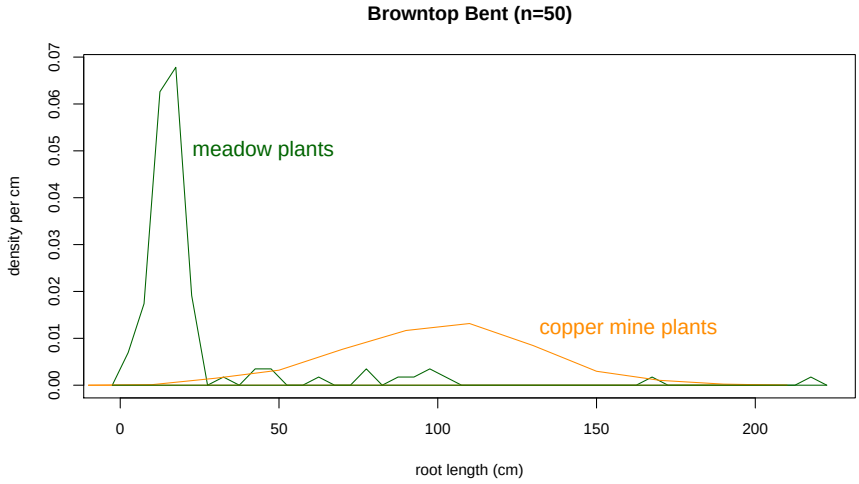
[Bra60] A.D. Bradshaw. Population Differentiation in *agrostis tenius Sibth.* III. populations in varied environments. *New Phytologist*, 59(1):92 – 103, 1960.

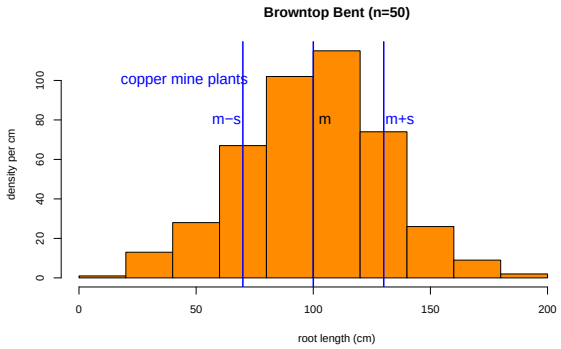
[MB68] T. McNeilly and A.D Bradshaw. Evolutionary Processes in Populations of Copper Tolerant Agrostis tenuis Sibth. *Evolution*, 22:108–118, 1968.

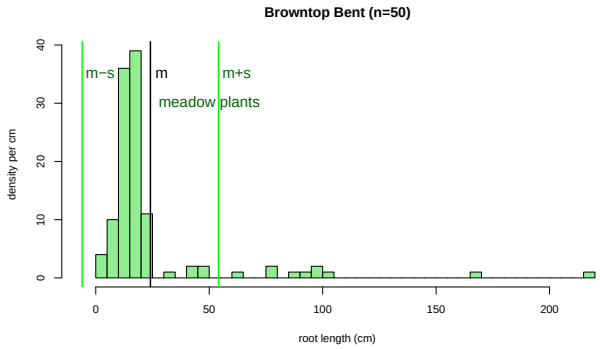
Wir verwenden hier wieder simulierte Daten, da die Originaldaten nicht zur Verfügung stehen.

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.





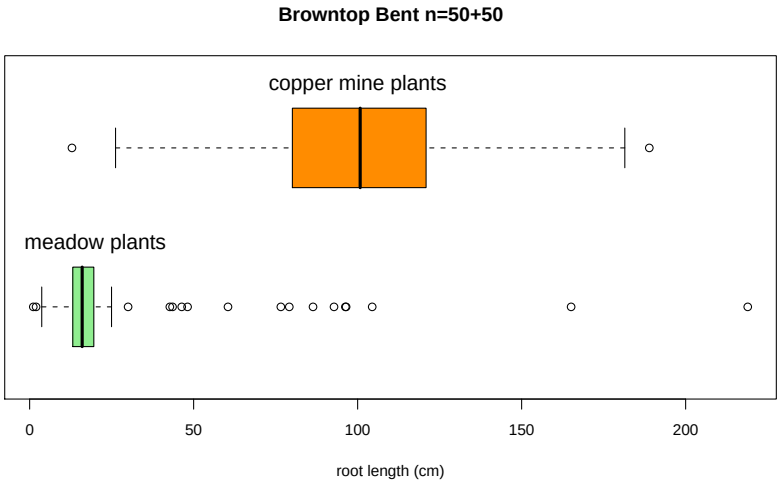


2/3 der Wurzellängen innerhalb [m-sd,m+sd]???? Nein!

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!