

Biostatistik

4. Deskriptive Statistik

Matthias Birkner



JOHANNES GUTENBERG
UNIVERSITÄT MAINZ

- 1 Wozu Statistik?
- 2 Graphische Darstellungen
 - Histogramme und Dichtepolygone
 - Stripcharts
 - Boxplots
 - Beispiel: Ringeltaube
 - Beispiel: Darwin-Finken
- 3 Statistische Kenngrößen
 - Median und andere Quartile
 - Mittelwert und Standardabweichung
- 4 Vom Sinn und Unsinn von Mittelwerten
 - Beispiel: Wählerische Bachstelzen
 - Beispiel: Spiderman & Spiderwoman
 - Beispiel: Kupfertoleranz beim Roten Straußgras
- 5 Das Statistikpaket R
 - Deskriptive Statistik mit R

*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

Katrin Kronenberger and Michael Türkay, A population study of *Galathea intermedia* in the German Bight, *Journal of the Marine Biological Association of the United Kingdom* 83:133–141, (2003).

M. Sonnewald, M. Apel, In Remembrance of Michael Türkay, a Tower of Strength in the World of Crustaceans, *Journal of Crustacean Biology* 36(1):106–117, (2016):

TAXA NAMED AFTER
PROF. DR. MICHAEL TÜRKAY (20)

- Ambilimbus tuerkayi* Martinez Arbizu, 1999 (Erebonasteridae)
Amphicrossus tuerkayi Martinez Arbizu, 1999 (Erebonasteridae)
Atyaephyra tuerkayi Cristodoulou, Antoniou, Magoulas and Koukouras, 2012 (Atyidae)
Bathymodiolus tangaroana tuerkayi von Cosel and Jannsen, 2008 (Mollusca: Mytilidae)
Calappa tuerkayana Pastore, 1995 (Calappidae)
Calyptogena tuerkayi Krylova and Jannsen, 2006 (Mollusca: Vesicomysidae)
Discoplax michalis Ng and Shih, 2015 (Gecarcinidae)
Divacuma tuerkayi Muhlenhardt-Siegel, 2003 (Diastylidae)
Euchirograpsus tuerkayi Crosnier, 2001 (Plagusiidae)
Eunoe tuerkayi Barnich and Fiege, 2003 (Polychaeta: Polynoidae)
Hymenopenaeus tuerkayi Crosnier, 1995 (Solenoceridae)
Lithodes turkayi Macpherson, 1987 (Lithodidae)
Mantisgebia tuerkayi Sakai, 2011 (Upogebiidae)
Nematocarcinus tuerkayi Burukovsky, 2005 (Nematocarcinidae)
Obliquogobius turkayi Goren, 1992 (Pisces: Gobiidae)
Petrolisthes tuerkayi Naderloo and Apel, 2014 (Porcellanidae)
Pterochirella tuerkayi Schulz, 1990 (Aetideidae)
Solitariopagurus tuerkayi McLaughlin, 1997 (Paguridae)
Tuerkayogebia Sakai, 1982 (Upogebiidae)
Tuerkayogebia kiiensis (Sakai, 1971) (Upogebiidae)



Eckart Witzigmann, Christian Teubner und Michael Türkay, *Shrimps, Hummer & Langusten. Warenkunde und Küchenpraxis der Krustentiere*, Gräfe und Unzer, 1994.

Der Springkrebs

Galathea intermedia



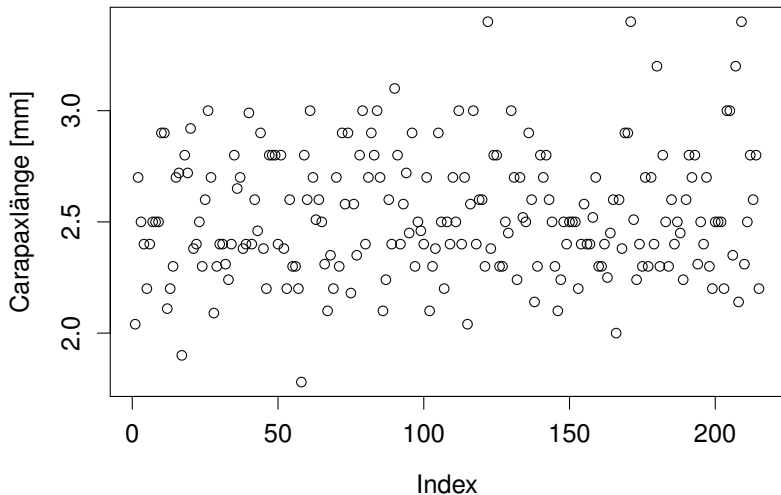
Helgoländer Tiefe Rinne, Fang vom 6.9.1988

Carapaxlänge (mm):

Nichteiertragende Weibchen ($n = 215$)

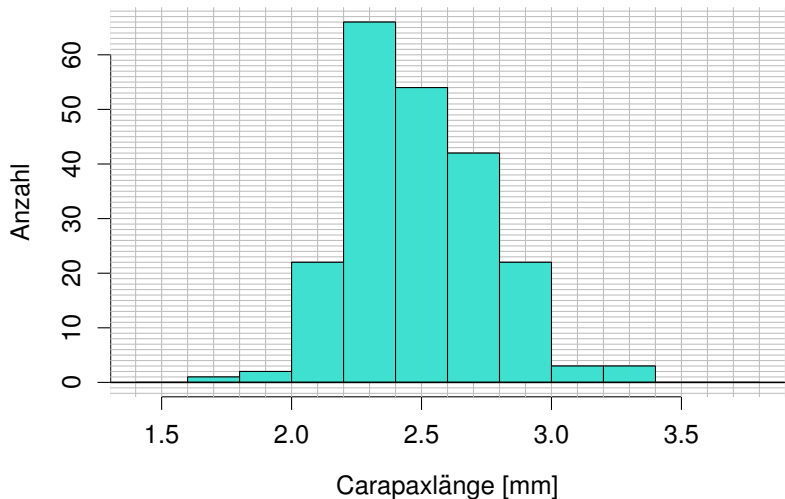
2,9	3,0	2,9	2,5	2,7	2,9	2,9	3,0
3,0	2,9	3,4	2,8	2,9	2,8	2,8	2,4
2,8	2,5	2,7	3,0	2,9	3,2	3,1	3,0
2,7	2,5	3,0	2,8	2,8	2,8	2,7	3,0
2,6	3,0	2,9	2,8	2,9	2,9	2,3	2,7
2,6	2,7	2,5	.	.	.	.	.

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

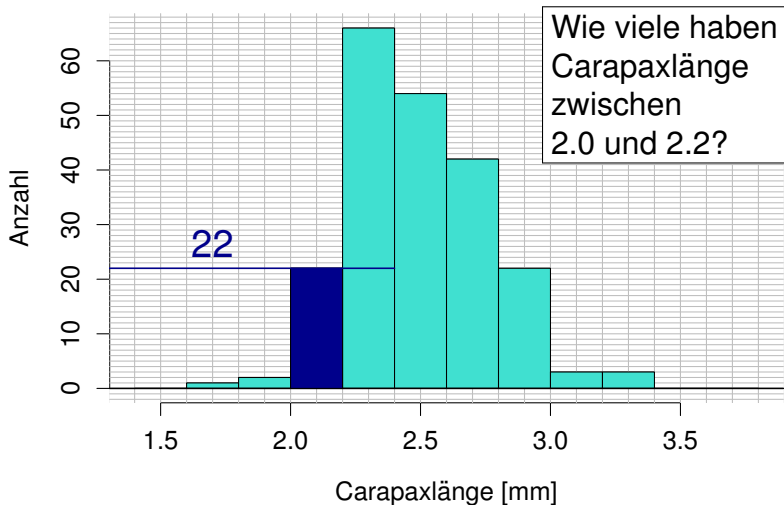


Eine Möglichkeit der graphischen
Darstellung:
das Histogramm

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

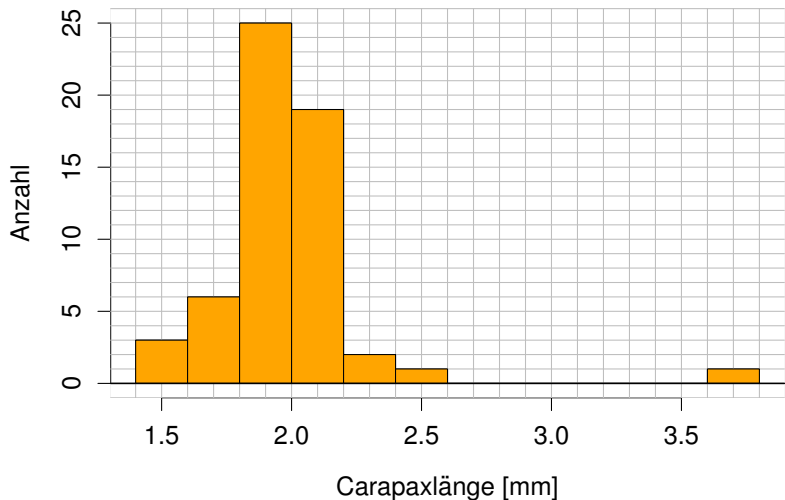


Nichteiertragende Weibchen am 6. Sept. '88, n=215



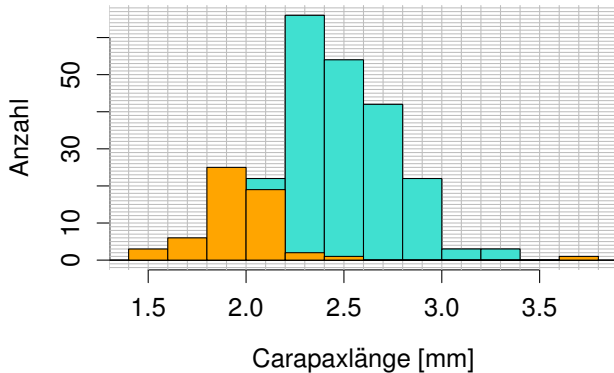
Analoge Daten zwei Monate später
(3.11.88):

Nichteiertragende Weibchen am 3. Nov. '88, n=57



Vergleich der beiden Verteilungen

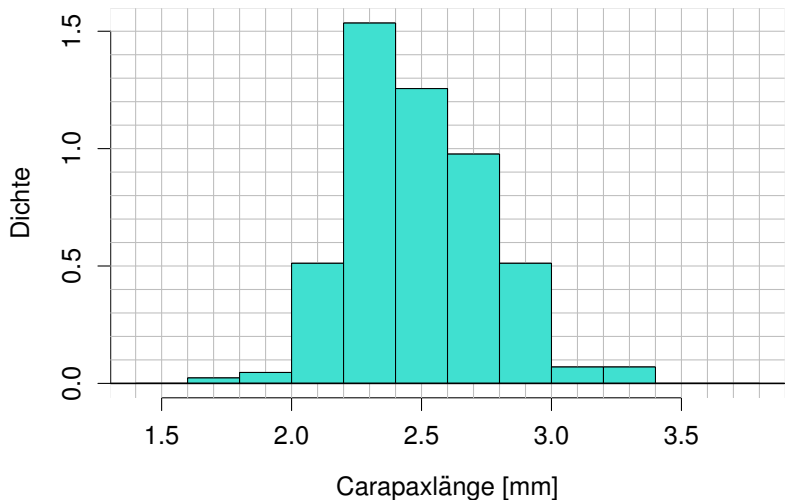
Nichteiertragende Weibchen



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

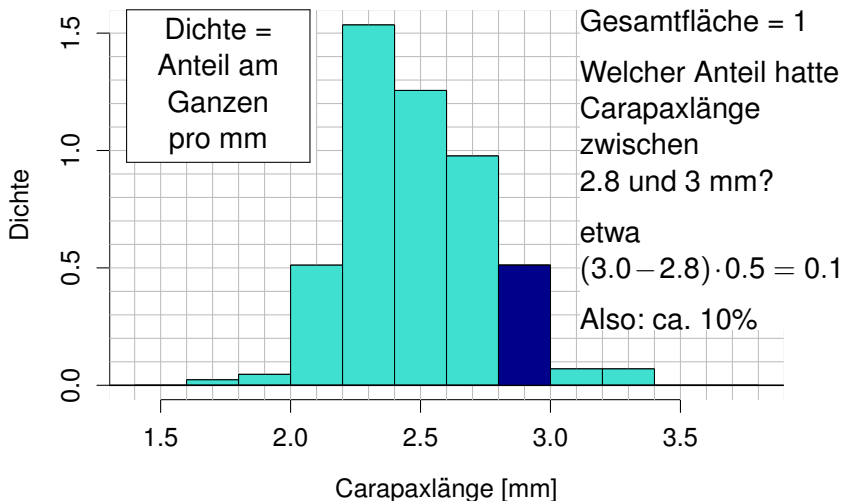
Idee: stauche vertikale Achse so, dass Gesamtfläche = 1.

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



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

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

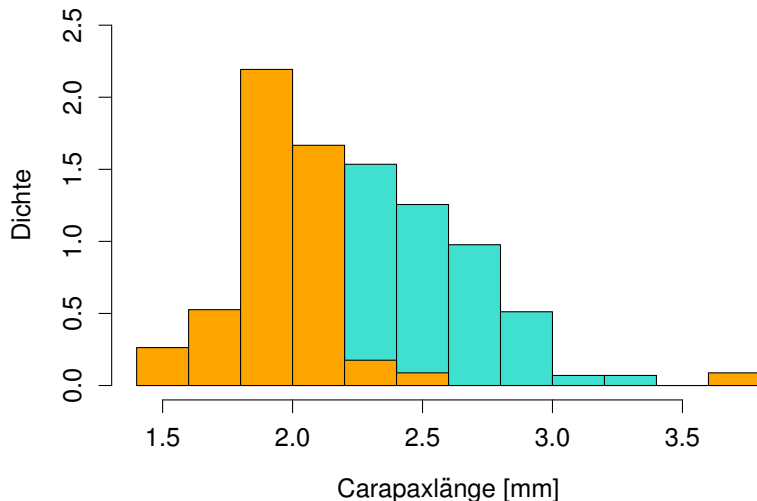


(Ganz penibel: Ich schreibe „ca. 10%“, denn die exakte Balkenhöhe ist 0.5116..., aber so genau wollen wir es hier gar nicht angeben.)

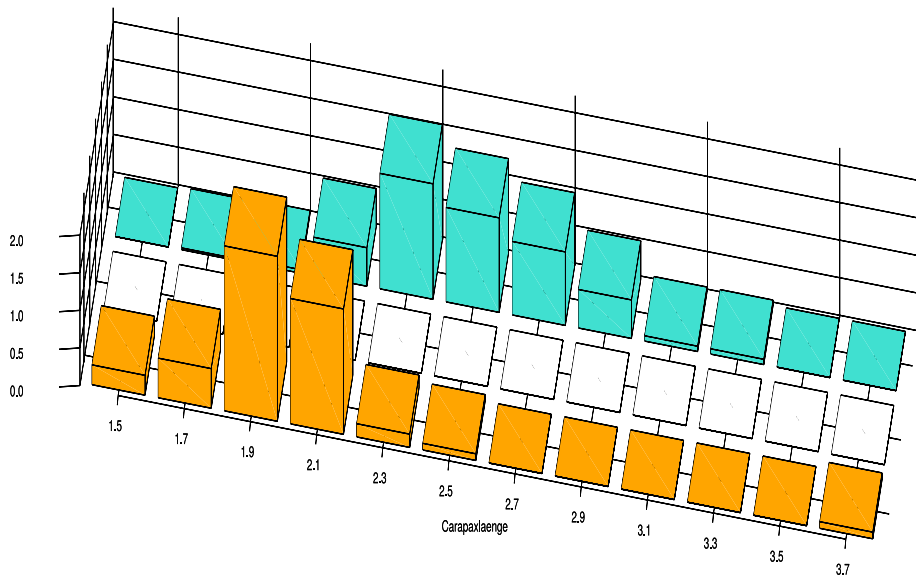
Die beiden Histogramme sind jetzt
vergleichbar
(sie haben dieselbe Gesamtfläche).

Versuche, die Histogramme zusammen zu zeigen:

Nichteiertragende Weibchen



Versuche, die Histogramme zusammen zu zeigen:



Vorschlag

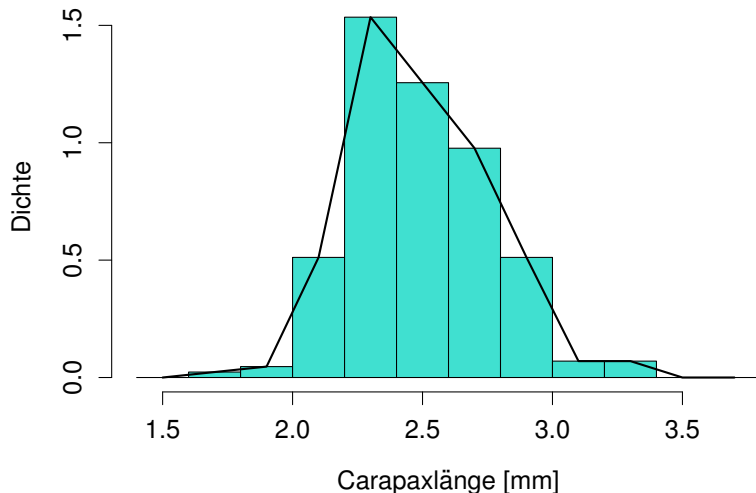
Total abgefahrene 3D-Plots
können in der Werbung nützlich sein,
für die Wissenschaft sind einfache und klare
2D-Darstellungen meistens angemessener.

Problem

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

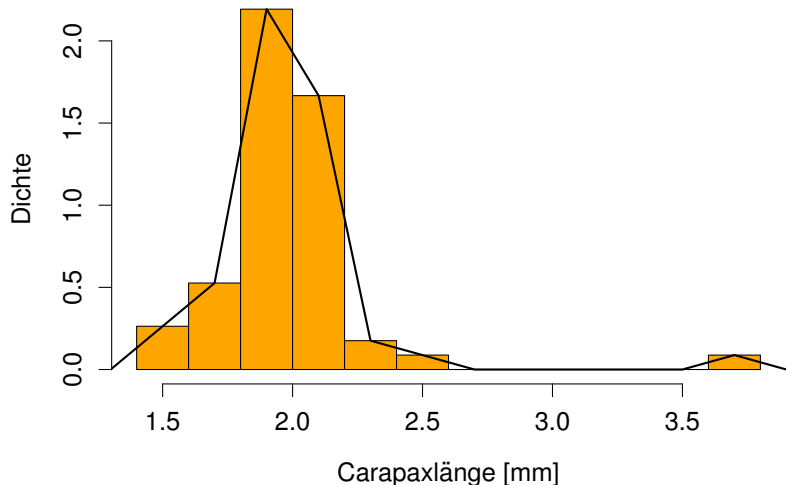
Einfache und klare Lösung: Dichtepolygone

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



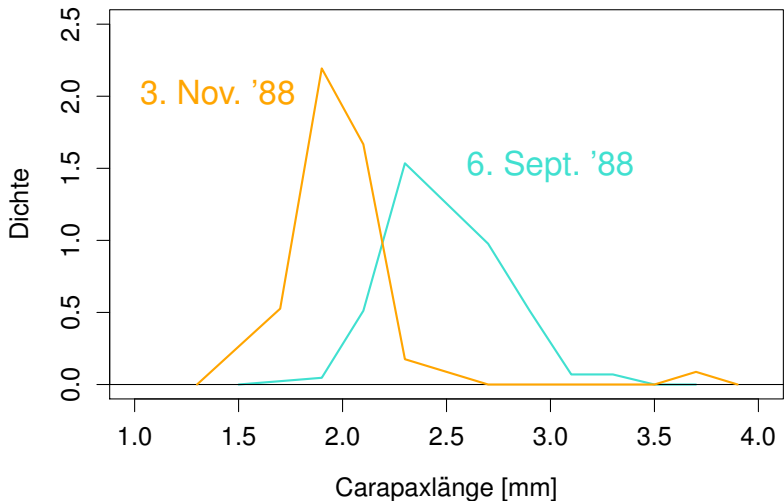
Einfache und klare Lösung: Dichtepolygone

Nichteiertragende Weibchen am 3. Nov. '88, n=57



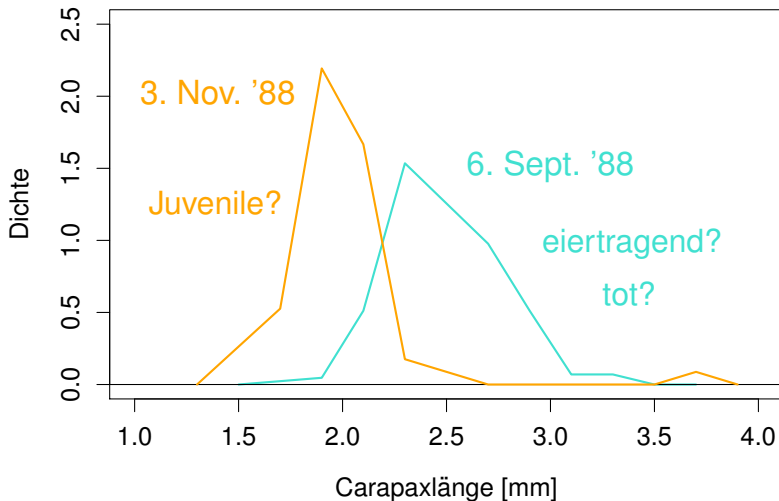
Zwei und mehr Dichtepolygone in einem Plot

Nichteiertragende Weibchen

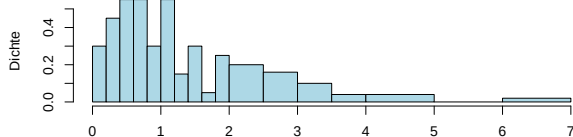
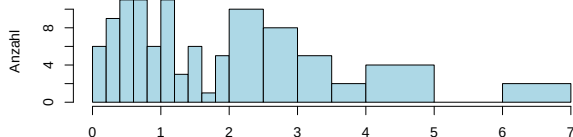
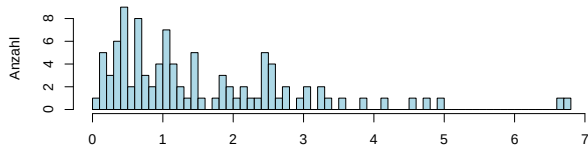


Biologische Interpretation der Verschiebung?

Nichteiertragende Weibchen



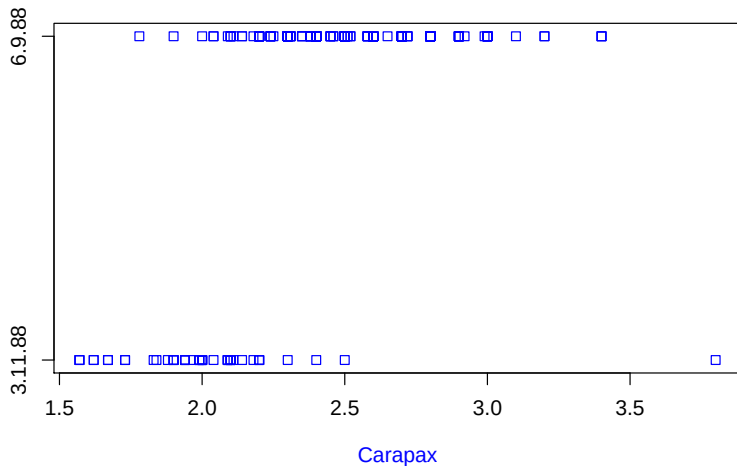
Anzahl vs. Dichte



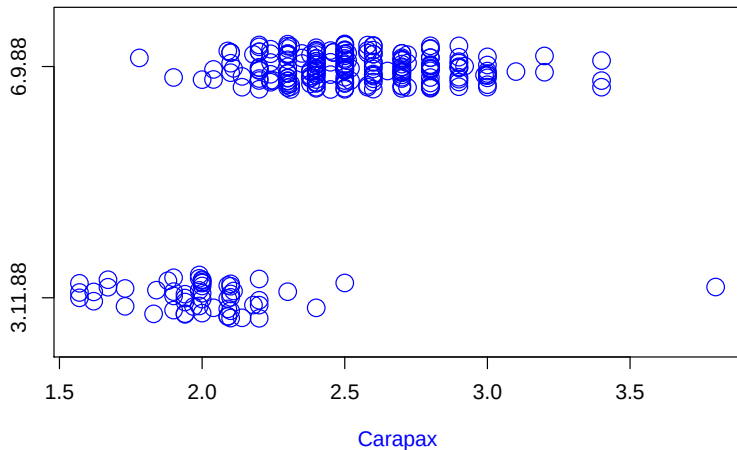
Also:

Bei Histogrammen
mit ungleichmäßiger
Unterteilung immer
Dichten verwenden!

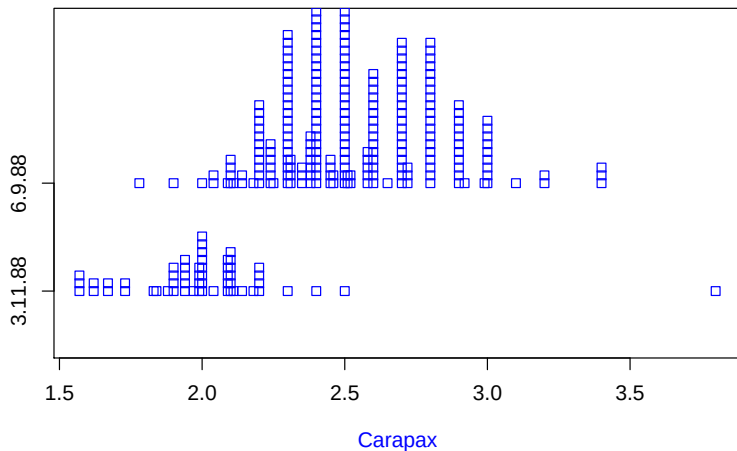
Einfacher Stripchart



Stripchart mit “jitter”



Stripchart mit “stacking”



Histogramme und Dichtepolygone
geben
ein ausführliches Bild
eines Datensatzes.

Manchmal zu ausführlich.

Zu viel Information erschwert den Überblick

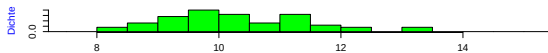


Baum Baum Baum Baum Baum Baum Baum Baum Baum Baum Baum Baum Baum Baum Baum

Wald?

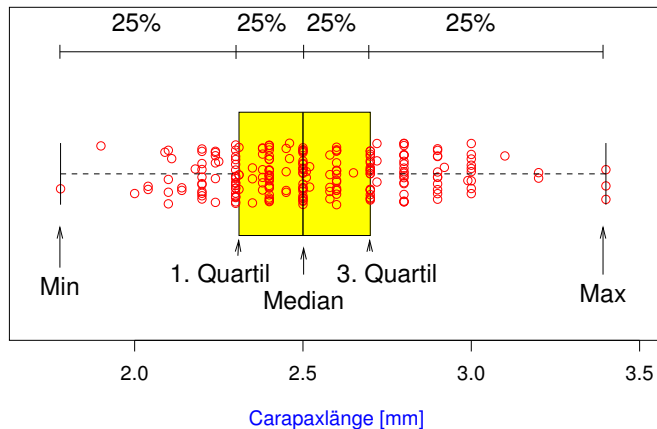
Beispiel:

Vergleich von mehreren Gruppen



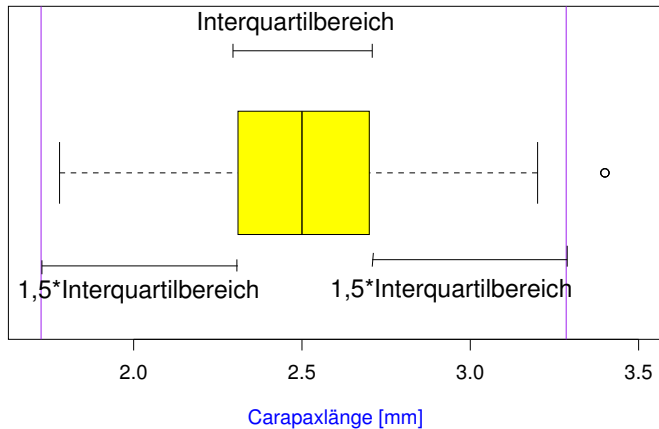
Der Boxplot

Boxplot, einfache Ausführung

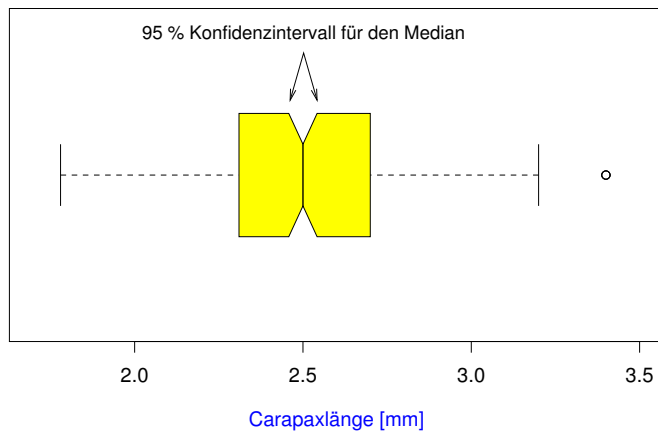


Der Boxplot

Boxplot, Standardausführung



Boxplot, Profiausstattung



Beispiel:

Die Ringeltaube

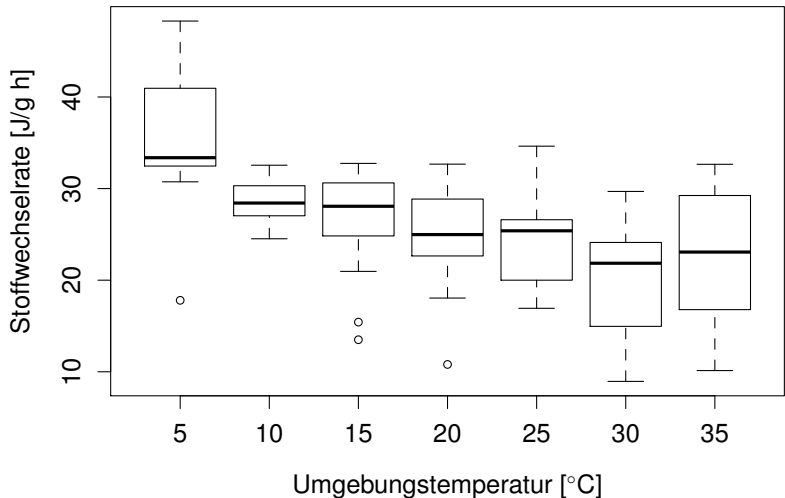
Palumbus palumbus



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

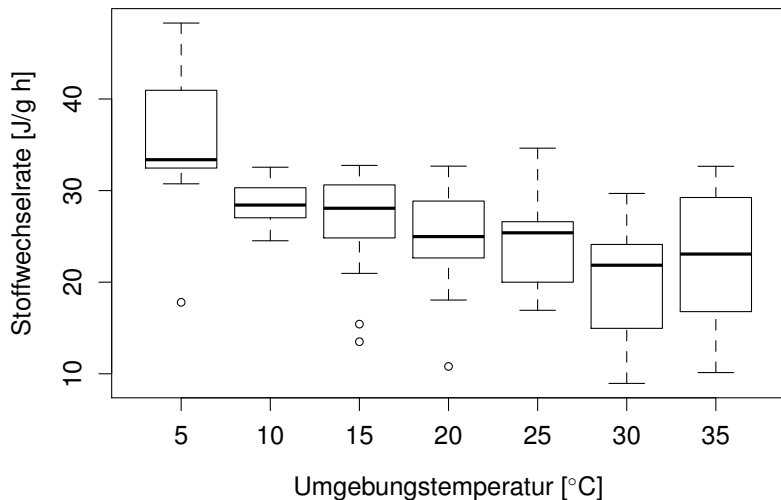
Daten aus dem AK Stoffwechselphysiologie,
Prof. Prinzinger, Universität Frankfurt

Stoffwechselrate und Temperatur bei Ringeltauben, n=90



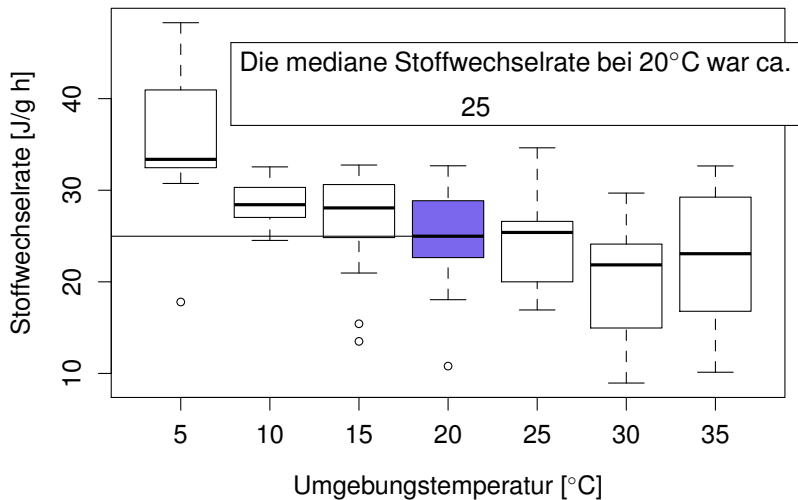
Klar:
Stoffwechselrate
höher
bei
tiefen Temperaturen

Stoffwechselrate und Temperatur bei Ringeltauben, n=90

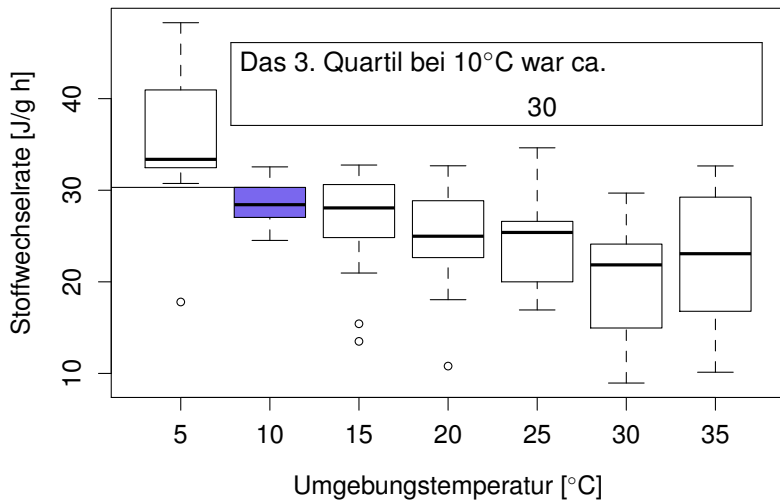


Vermutung:
Bei **hohen** Temperaturen
nimmt die Stoffwechselrate
wieder zu
(Hitzestress).

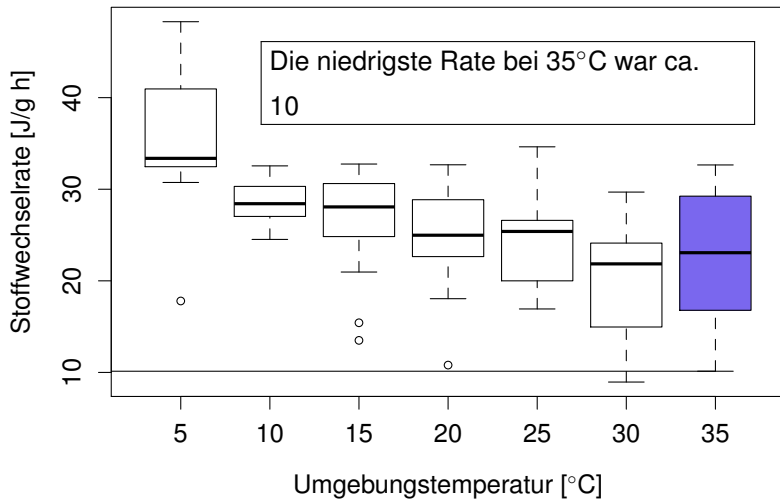
Stoffwechselrate und Temperatur bei Ringeltauben, n=90



Stoffwechselrate und Temperatur bei Ringeltauben, n=90



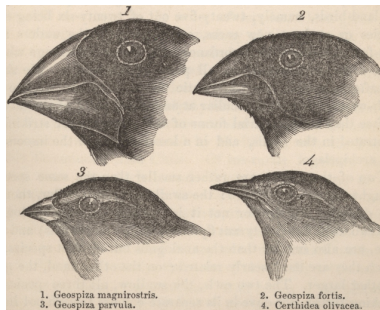
Stoffwechselrate und Temperatur bei Ringeltauben, n=90



Charles Robert Darwin (1809-1882)



Darwin-Finken



[http:](http://darwin-online.org.uk/graphics/Zoology_Illustrations.html)

[//darwin-online.org.uk/graphics/Zoology_Illustrations.html](http://darwin-online.org.uk/graphics/Zoology_Illustrations.html)

Darwins Finken-Sammlung

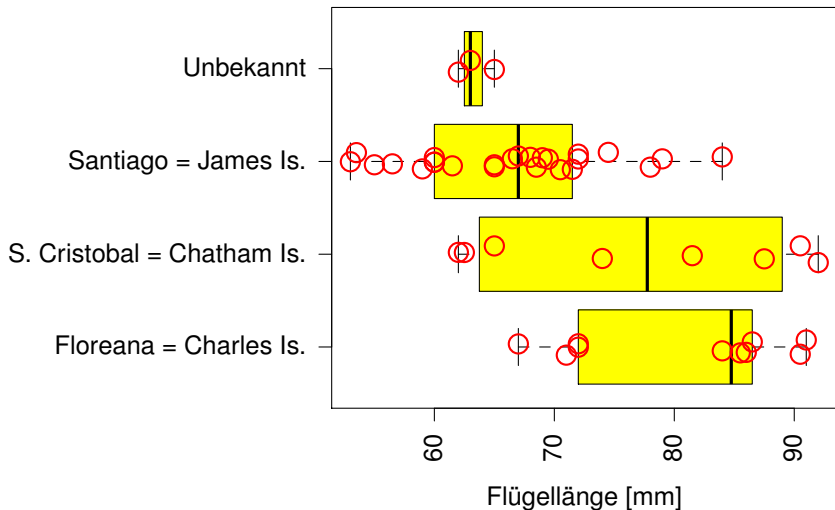


Sulloway, F.J. (1982) The Beagle collections of Darwin's Finches (Geospizinae). *Bulletin of the British Museum (Natural History), Zoology series* **43**: 49-94.

- ▶ <https://datadryad.org/stash/dataset/doi:10.5061/dryad.154>

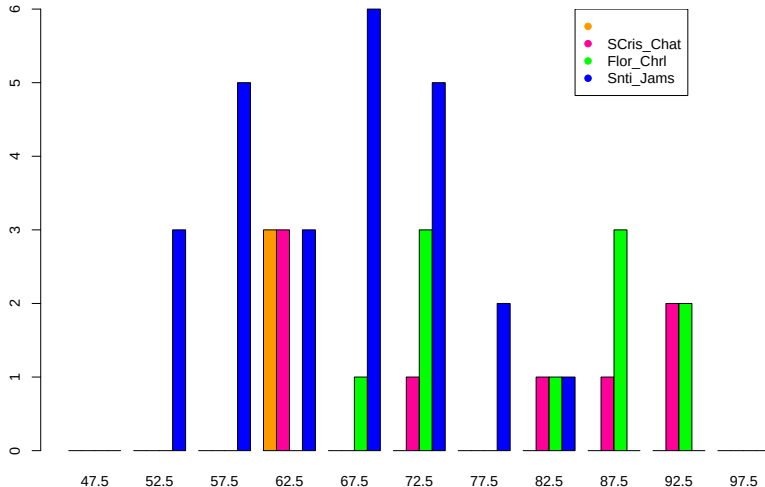
Flügelängen der Darwin-Finken

Flügelängen nach Insel (ges. n=49)

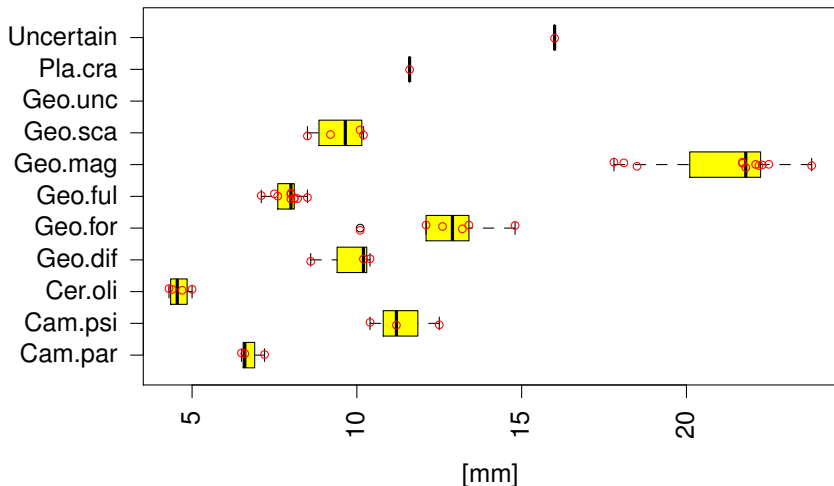


Flügelängen der Darwin-Finken

Barplot für Flügelängen (Anzahlen)



Schnabelgröße je nach Art



(Cam.par=*Camarhynchus parvulus*, Cam.psi=*Camarhynchus psittacula*, Cer.oli=*Certhidea olivacea*, Geo.dif=*Geospiza difficilis*, Geo.for=*Geospiza fortis*, Geo.ful=*Geospiza fuliginosa*, Geo.mag=*Geospiza magnirostris*, Geo.sca=*Geospiza scandens*, Geo.unc=*Geospiza (unklar)*, Pla.cra=*Platypiza crassirostris*)

Fazit

- 1 Histogramme erlauben einen detaillierten Blick auf die Daten
- 2 Dichtepolygone erlauben Vergleiche zwischen vielen Verteilungen
- 3 Boxplot können große Datenmengen vereinfacht zusammenfassen
- 4 Bei kleinen Datenmengen eher Stripcharts 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.

* Diese „Definition“ genügt für die meisten praktischen Fälle (und ist intuitiv sehr plausibel), die mathematisch präzise Definition siehe die folgende Folie.

Nachtrag:

Der Median:

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

Der Median ist das 50%-Quantil der Daten.

*Eine mathematisch präzise Definition:

Seien n der Größe nach geordnete Beobachtungswerte $y_1 \leq y_2 \leq \dots \leq y_n$ gegeben, dann ist (der/ein) Median m ein Wert, so dass höchstens $n/2$ Werte $\geq m$ und höchstens $n/2$ Werte $\leq m$ sind.

Falls n ungerade ist, sagen wir $n = 2k + 1$, so ist durch diese Forderung $m = y_{k+1}$ eindeutig festgelegt, denn

$$\underbrace{y_1, y_2, \dots, y_k}_{k \text{ Werte}} \leq y_{k+1} \leq \underbrace{y_{k+2}, y_{k+3}, \dots, y_{2k+1}}_{k \text{ Werte}}.$$

Falls n gerade ist, sagen wir $n = 2k$, so erfüllen y_k, y_{k+1} und ggfs. auch jeder Wert zwischen y_k und y_{k+1} diese Forderung. Wenn ein konkreter Wert verlangt wird, nimmt man dann oft pragmatisch $(y_{k+1} + y_k)/2$ — so ist es beispielsweise in R implementiert.

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.

*

Auch hier: Diese „Definition“ genügt für die meisten praktischen Fälle (und ist intuitiv sehr plausibel), für eine mathematisch präzise Definition siehe die übernächste Folie.

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.

*

Auch hier: Diese „Definition“ genügt für die meisten praktischen Fälle (und ist intuitiv sehr plausibel), für eine mathematisch präzise Definition siehe die folgende Folie.

Nachtrag:

Erstes Quartil, Q_1 : ein Viertel der Beobachtungen sind kleiner, drei Viertel sind größer*.

Drittes Quartil, Q_3 : drei Viertel der Beobachtungen sind kleiner, ein Viertel sind größer†.

*Präziser kann man Folgendes fordern: Q_1 ist eine Zahl, so dass

- höchstens 25% der Beobachtungswerte $< Q_1$ und
- höchstens 75% der Beobachtungswerte $> Q_1$ sind.

†Präziser kann man Folgendes fordern: Q_3 ist eine Zahl, so dass

- höchstens 75% der Beobachtungswerte $< Q_3$ und
- höchstens 25% der Beobachtungswerte $> Q_3$ sind.

Ähnlich wie beim Median kann es hier prinzipiell vorkommen, dass verschiedene Zahlen in Frage kommen, die alle diese Bedingungen erfüllen. In der Literatur gibt es verschiedene Konventionen, welche man dann genau nehmen sollte (sie beispielsweise die Online-Hilfe von R zum Befehl `quantile`), in den meisten „realistischen“ Datensätzen unterscheiden sich die Antworten aber kaum.

Am häufigsten werden benutzt:

Lageparameter

Der Mittelwert $\bar{x}$

Streuungsparameter

Die Standardabweichung s

Der Mittelwert

(engl. *mean*)

Notation:

Wenn die Beobachtungen

$$x_1, x_2, x_3, \dots, x_n$$

heißen,

schreibt man oft

$$\bar{x}$$

für den Mittelwert.

Definition:

Mittelwert

=

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

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

Beispiel:

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

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

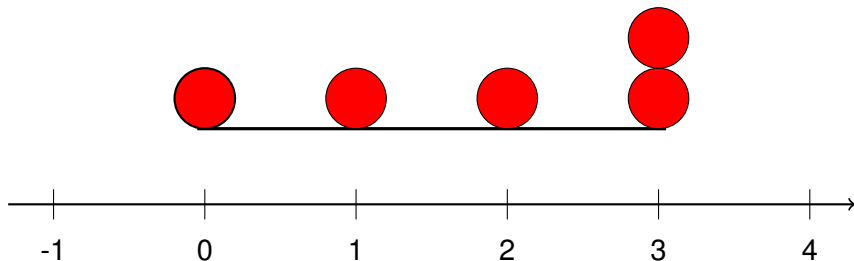
$$\bar{x} = (3 + 0 + 2 + 3 + 1) / 5$$

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

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

Geometrische Bedeutung des Mittelwerts: Der Schwerpunkt

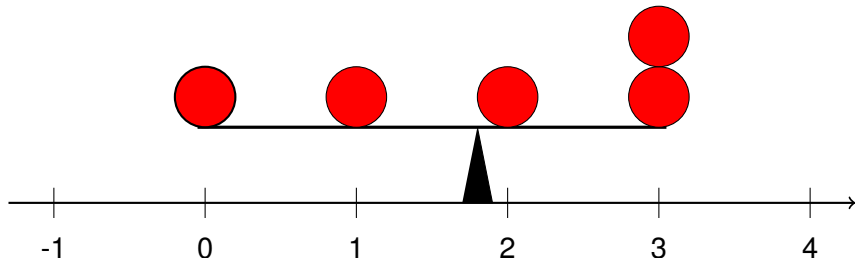
Wir stellen uns die Beobachtungen
als gleich schwere Gewichte
auf einer Waage vor:



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

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

$$\bar{x} = 2 \quad ? \quad \bar{x} = 1.5 \quad ? \quad \bar{x} = 1.8 \quad ?$$



richtig!

Beispiel: *Galathea intermedia*

„Rundlichkeit“

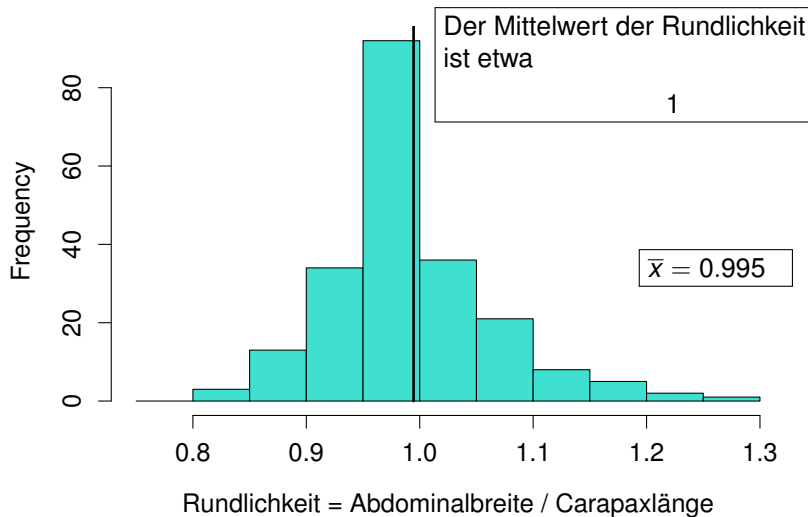
$\hat{=}$

Abdominalbreite / Carapaxlänge

Vermutung:

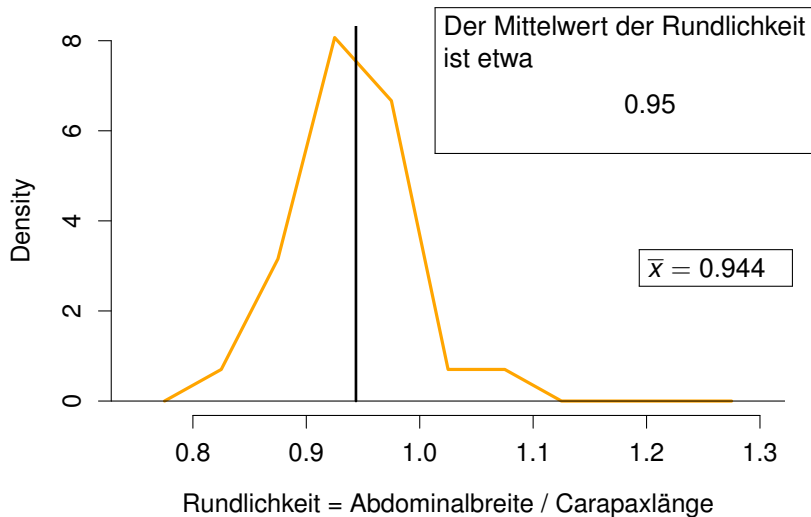
Rundlichkeit nimmt
bei Geschlechtsreife zu

Nichteiertragende Weibchen am 6. Sept. '88



Noch ein Beispiel:
Die Daten vom 3.11.88

Nichteiertragende Weibchen am 3. Nov. '88

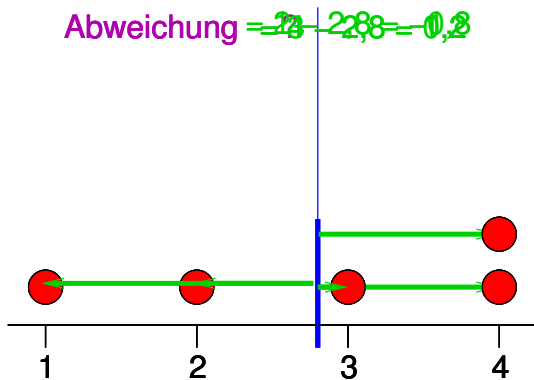


Die Standardabweichung

Wie weit weicht
eine typische Beobachtung
vom
Mittelwert
ab ?

typische Mittelwert=2,8

$$\text{Abweichung} = 2,8 - 2,8 = 0,2$$



Die **Standardabweichung** σ (“sigma”)

[auch *SD* von engl. *standard deviation*]

ist ein

etwas komisches

gewichtetes Mittel

der Abweichungsbeträge

und zwar

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

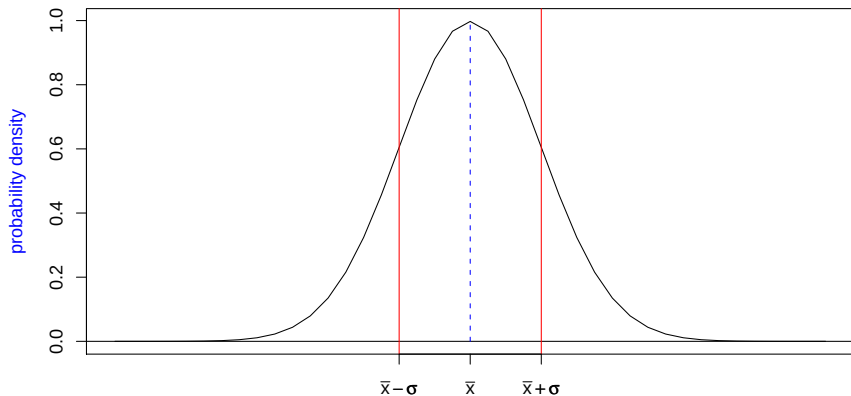
Die **Standardabweichung** von $x_1, x_2, \dots, x_n$
als Formel:

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

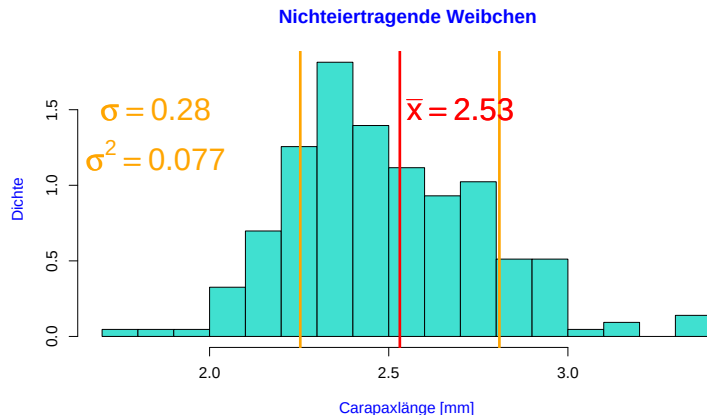
$$\sigma^2 = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2 \text{ hei\ss}t \text{ **Varianz**.}$$

Faustregel für die Standardabweichung

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



Standardabweichung der Carapaxlängen nichteiertragender Weibchen vom 6.9.88



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

Varianz der Carapaxlängen nichtteiertragender 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:

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

Können wir 0,0768 als Schätzwert für die Varianz $\sigma_{\mathcal{X}}^2$ in der ganzen Population verwenden?

Ja, können wir machen. Allerdings ist σ_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: $\sigma_X^2 = \frac{1}{N} \sum_{i=1}^N (X_i - \bar{X})^2$

Stichprobenvarianz: $\sigma_S^2 = \frac{1}{n} \sum_{i=1}^n (S_i - \bar{S})^2$

korrigierte Stichprobenvarianz:

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

Mit “Standardabweichung von S ” ist
(für Daten oder Stichproben) meistens das korrigierte s gemeint.

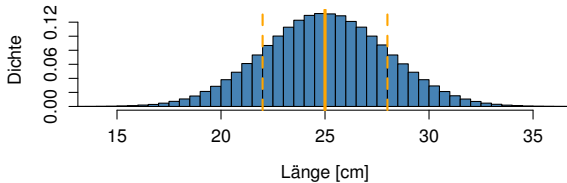
$$\bar{x} = 10/5 = 2$$

Summe

i	1	2	3	4	5	
x_i	1	3	0	5	1	10
$x_i - \bar{x}$	-1	1	-2	3	-1	0
$(x_i - \bar{x})^2$	1	1	4	9	1	16

$$s^2 = \frac{1}{5-1} \sum_{i=1}^5 (x_i - \bar{x})^2 = \frac{16}{5-1} = 4$$

$$s = 2$$

Eine simulierte Fischpopulation ($N = 10^6$ Adulti)

Mittlere Länge in der Population:

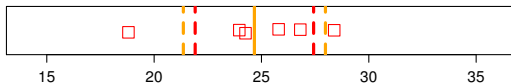
$$\mu = 25 \text{ [cm]}$$

Standardabweichung in der Population:

$$\sigma = 3 \text{ [cm]}$$

d.h. Varianz

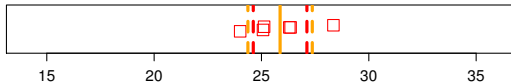
$$\sigma^2 = 9 \text{ [cm}^2\text{]}$$

Eine Stichprobe ($n = 6$)

Hier: $\bar{x} = 24.67$

SD mit $(n - 1)$: 3.31

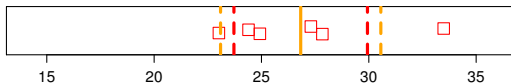
SD mit n : 2.76

Eine weitere Stichprobe ($n = 6$)

Hier: $\bar{x} = 25.87$

SD mit $(n - 1)$: 1.5

SD mit n : 1.25

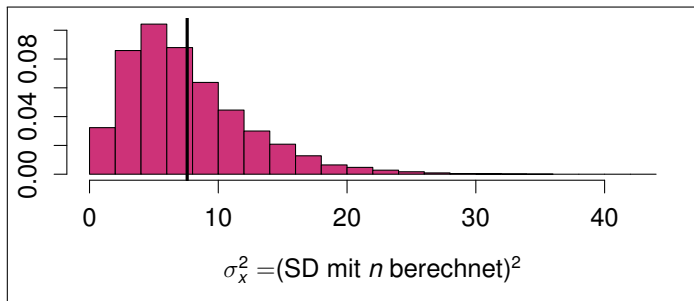
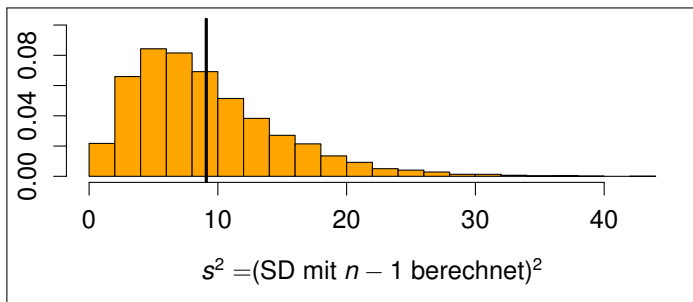
Eine dritte Stichprobe ($n = 6$)

Hier: $\bar{x} = 26.83$

SD mit $(n - 1)$: 3.73

SD mit n : 3.11

Bei 10000 simulierten Stichproben der Größe $n = 6$ finden wir:



σ versus s : mit n oder $n - 1$ berechnen?

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

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

Wenn es sich bei $x_1, \dots, x_n$ um eine Stichprobe aus einer großen „Population“ handelt (wie meistens in der Statistik), sollten Sie die Formel

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

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 simulierte Daten, wenn die Originaldaten nicht verfügbar waren.
(Nehmen Sie also nicht alle Datenpunkte wörtlich.)

Bachstelzen fressen Dungfliegen

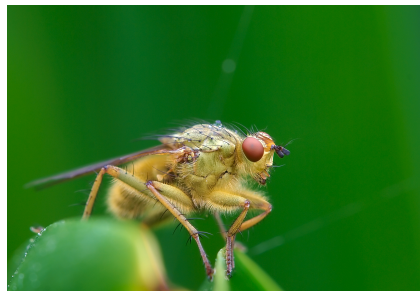
Räuber



Bachstelze (White Wagtail)
Motacilla alba alba

image (c) by Artur Mikołajewski

Beute



Gelbe Dungfliege
Scatophaga stercoraria

image (c) by Viatour Luc

Vermutung

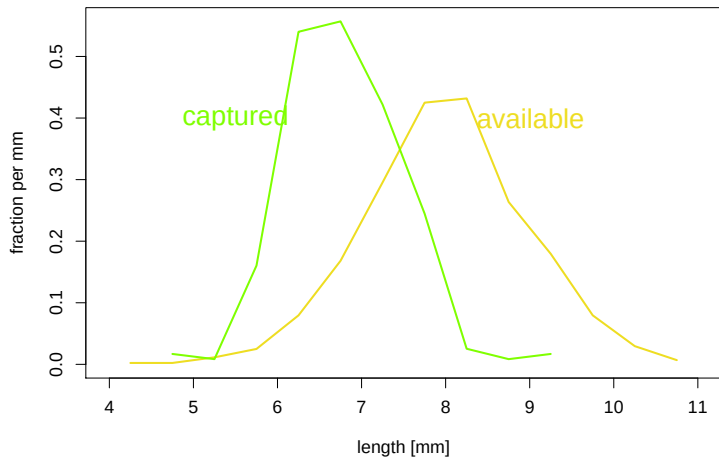
- Die Fliegen sind unterschiedlich groß
- Effizienz für die Bachstelze = Energiegewinn / Zeit zum Fangen und fressen
- Laborexperimente lassen vermuten, dass die Effizienz bei 7mm großen Fliegen maximal ist.



N.B. Davies.

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

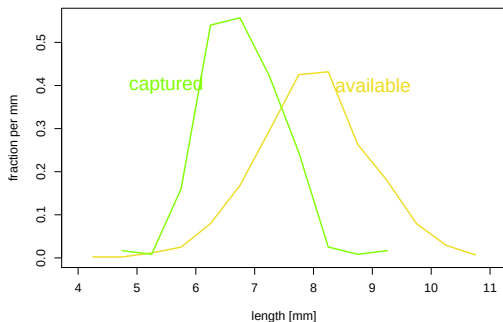
J. Anim. Ecol., 46:37–57, 1977.

dung flies: available, captured

Vergleich der Größenverteilungen

	captured		available
Mittelwert	6.79	<	7.99
Standardabweichung	0.69	<	0.96

dung flies: available, captured



Interpretation

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

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



Nephila madagascariensis

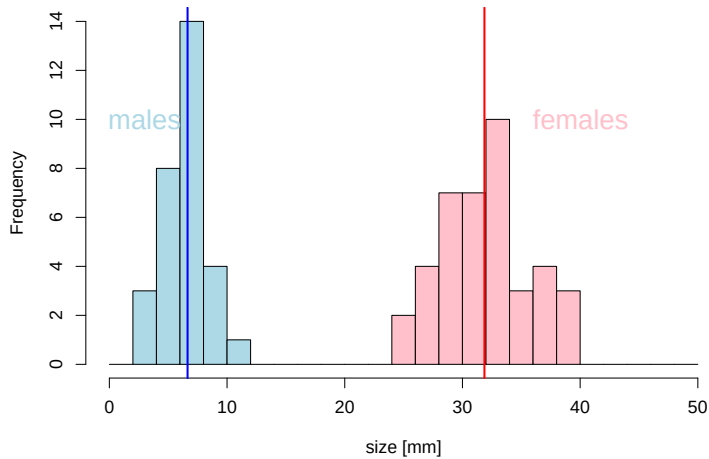
image (c) by Bernard Gagnon

Simulierte daten:

Eine Stichprobe von 70 Spinnen

Mittlere Größe: 21,06 mm

Standardabweichung der Größe: 12,94 mm

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



Nephila madagascariensis

image (c) by Arthur Chapman

Fazit des Spinnenbeispiels

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

Kupfertolerantes Rotes Straußgras



Rotes Straußgras
Agrostis tenuis

image (c) Kristian Peters



Kupfer
Cuprum

Hendrick met de Bles



A.D. Bradshaw.

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

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



T. McNeilly and A.D Bradshaw.

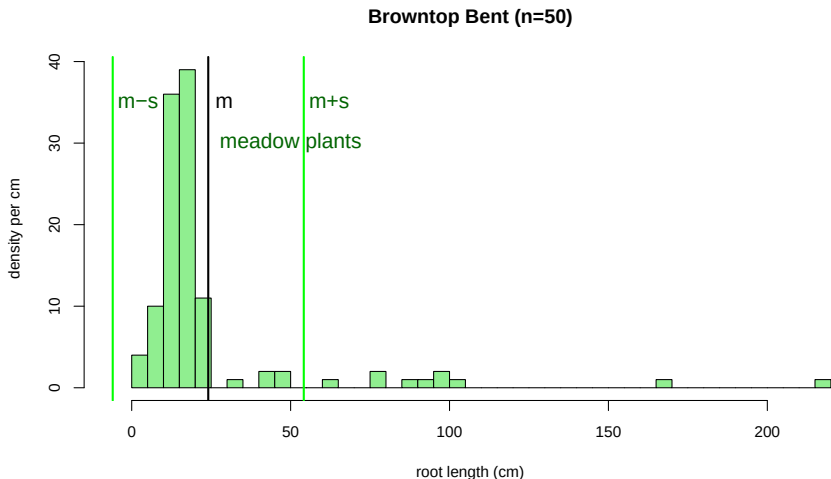
Evolutionary Processes in Populations of Copper Tolerant
Agrostis tenuis Sibth.

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 eingesät.
- Die Wurzellängen dieser “Wiesenpflanzen” werden gemessen.

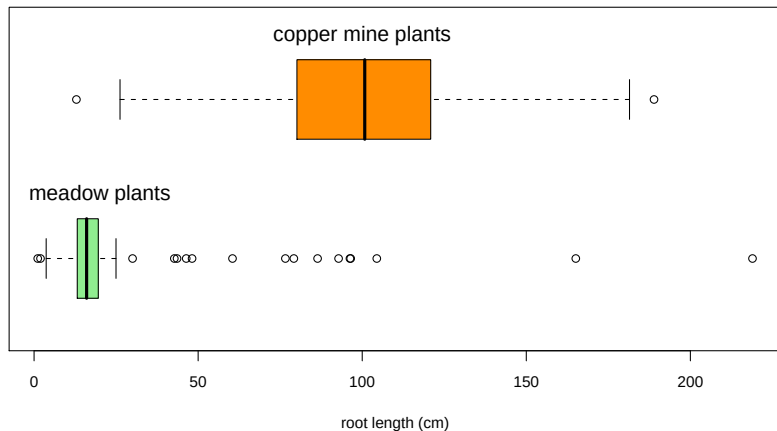


2/3 der Wurzellängen innerhalb $[m-sd, m+sd]$???? **Nein!**

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

Browntop Bent n=50+50

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!

Was ist R?

- Wir verwenden R in dieser Vorlesung als (sehr mächtigen) „Statistik-Taschenrechner.“
(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 hat eine sehr aktive Benutzer- und Entwicklergemeinde (die nahezu alle Bereiche der Statistik und viele Anwendungsbereiche 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.
- Im Browser auch via:
<https://jupyterhub.zdv.uni-mainz.de/>
(dort “RStudio environment” auswählen)

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: In der Übung ansprechen

R starten: Windows, Mac OS: Icon (ggf. aus Menu) anklicken,
Linux/Unix: `> R` auf einer Konsole

(Manche Benutzer finden auch die erweiterte graphische Oberfläche `rstudio` angenehm, <https://posit.co/downloads/>).

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

laufende Rechnungen unterbrechen: `CTRL-C`

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 wenige R-Befehle verwenden, so dass Sie i.A. keine über die Folien hinausgehende 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 Computing

www.r-project.org

Meistbesucht Getting Started Latest Headlines



About R
[What is R?](#)
[Contributors](#)
[Screenshots](#)
[What's new?](#)

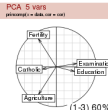
Download, Packages
[CRAN](#)

R Project
[Foundation](#)
[Members & Donors](#)
[Mailing Lists](#)
[Bug Tracking](#)
[Developer Page](#)
[Conferences](#)
[Search](#)

Documentation
[Manuals](#)
[FAQs](#)
[The R Journal](#)
[Wiki](#)
[Books](#)
[Certification](#)
[Other](#)

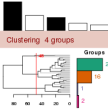
Misc
[Bioconductor](#)
[Related Projects](#)
[User Groups](#)
[Links](#)

PCA 5 vars
 principal(1) = 54%, (2) = 30%

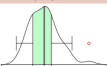


(1-3) 60%

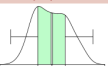
Clustering 4 groups



Factor 1 [41%]



Factor 3 [19%]



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 \(Wirtschaftsuniversität Wien\)](#).