

# Inhaltsverzeichnis

<b>1 Einführung</b>	<b>1</b>
1.1 Konzept und Quellen . . . . .	1
1.2 Plan . . . . .	2
<b>2 Ziele der deskriptiven (d.h. beschreibenden) Statistik</b>	<b>3</b>
<b>3 Graphische Darstellungen</b>	<b>3</b>
3.1 Histogramme und Dichtepolygone . . . . .	5
3.2 Stripcharts . . . . .	10
3.3 Boxplots . . . . .	11
3.4 Beispiel: Ringeltaube . . . . .	13
3.5 Beispiel: Darwin-Finken . . . . .	16
<b>4 Statistische Kenngrößen</b>	<b>18</b>
4.1 Median und andere Quartile . . . . .	19
4.2 Mittelwert und Standardabweichung . . . . .	19

## 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

#### Was ist Statistik?

Die Natur ist voller Variabilität.

Wie geht man mit variablen Daten um?

Es gibt eine mathematische Theorie des Zufalls:[0.5ex] die *Stochastik*.

#### IDEE DER STATISTIK

*Variabilität*

(Erscheinung der Natur)

durch

*Zufall*

(mathematische Abstraktion)

modellieren.

Statistik  
=  
*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

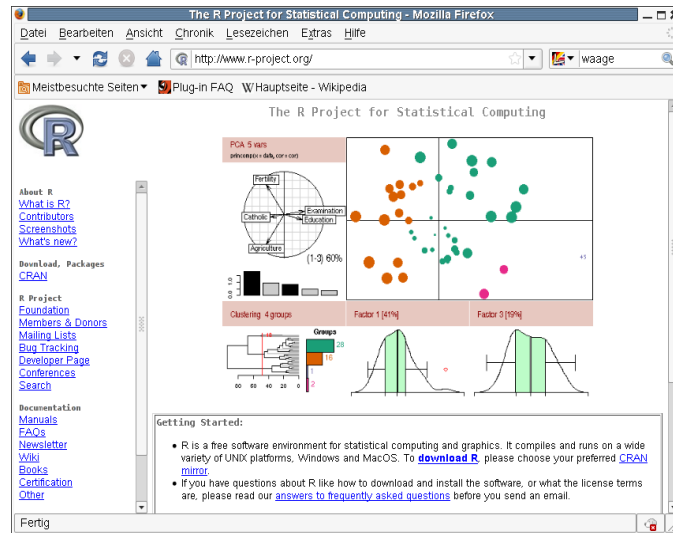
1. Beschreibende Statistik
2. Der Standardfehler
3. Der t-Test für gepaarte Stichproben
4. Der t-Test für unabhängige Stichproben
5. Häufigkeiten
6. Der Chi-Quadrat Test
7. Lineare Regression
8. Korrelation
9. Varianzanalyse (ANOVA)
10. Moderne Anwendung: Analyse von Genexpressionsdaten (vielleicht)
11. R

### Plan der Vorlesung

#### Weitere Themen

- Nichtparametrische Tests
- Diskriminanzanalyse
- Grundbegriffe der Wahrscheinlichkeitstheorie
- Parameterschätzung
- Moderne Anwendung: Analyse von Genexpressionsdaten

# Statistik-Software R



<http://www.r-project.org>

Folien, R-Befehle, Quellen und Übungen

<http://evol.bio.lmu.de/statgen/StatBiol/10SS>

## 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*

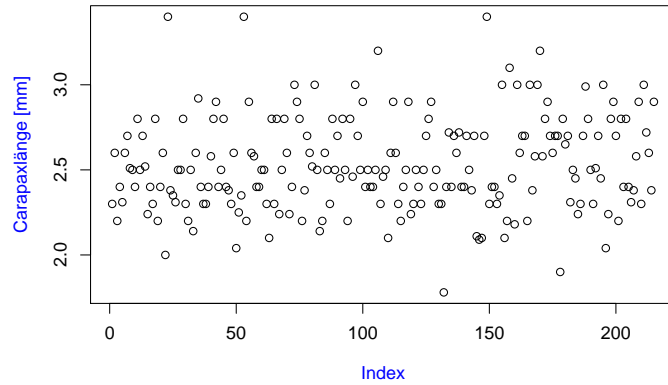


Helgoländer Tiefe Rinne, Fang vom 6.9.1988

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	.	.	.	.	.

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

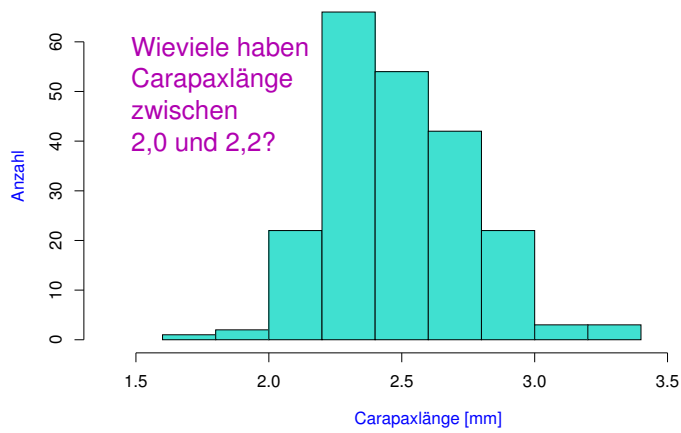


### 3.1 Histogramme und Dichtepolygone

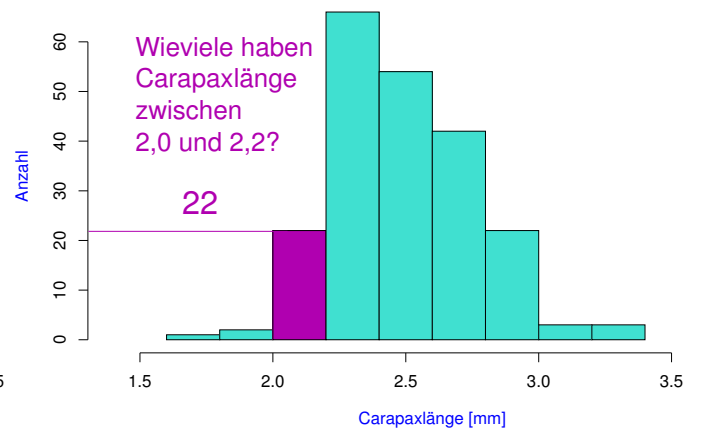
Eine Möglichkeit der graphischen Darstellung:

das Histogramm

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

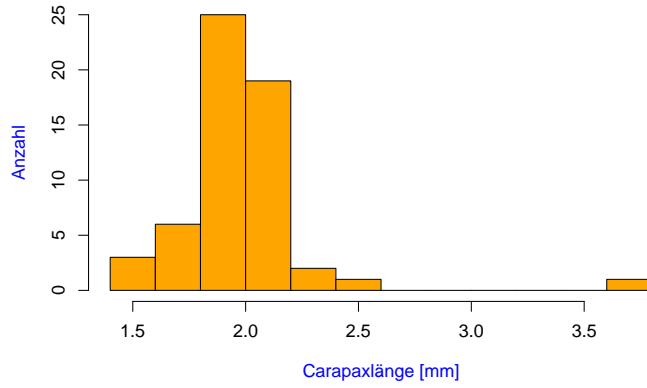


Nichteiertragende Weibchen am 6. Sept. '88, n=215

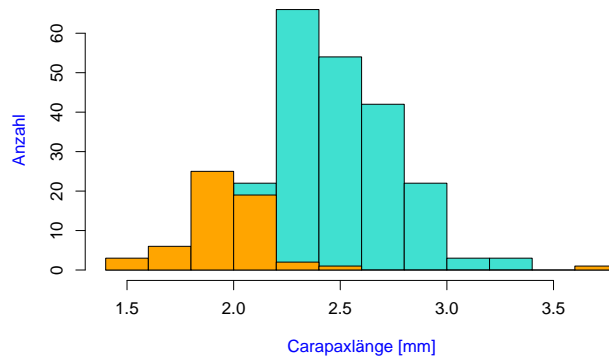


Analoge Daten zwei Monate später (3.11.88):

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



Nichteiertragende Weibchen

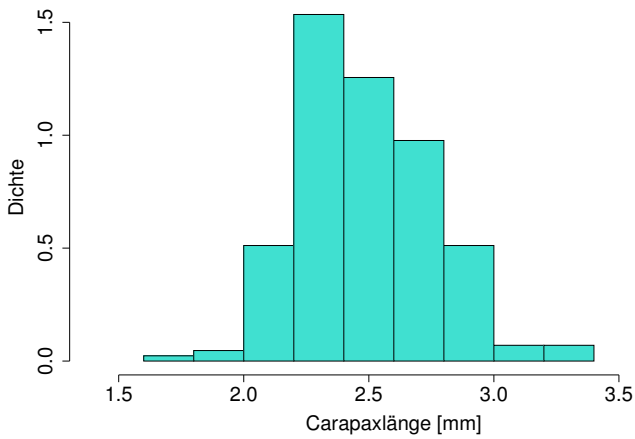


Vergleich der beiden Verteilungen

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

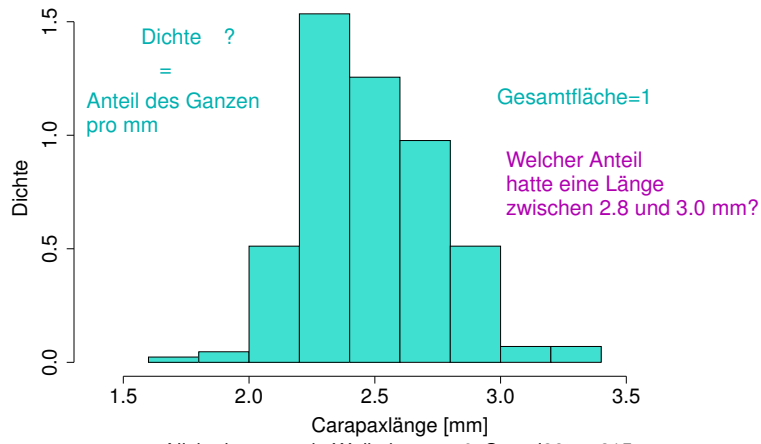
Idee: stauche vertikale Achse so, dass Gesamtfläche = 1.

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

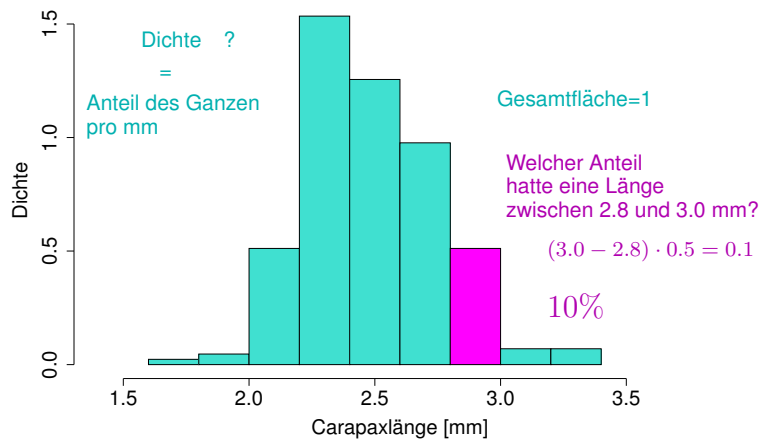


Die neue vertikale Koordinate ist jetzt eine *Dichte* (engl. *density*).

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

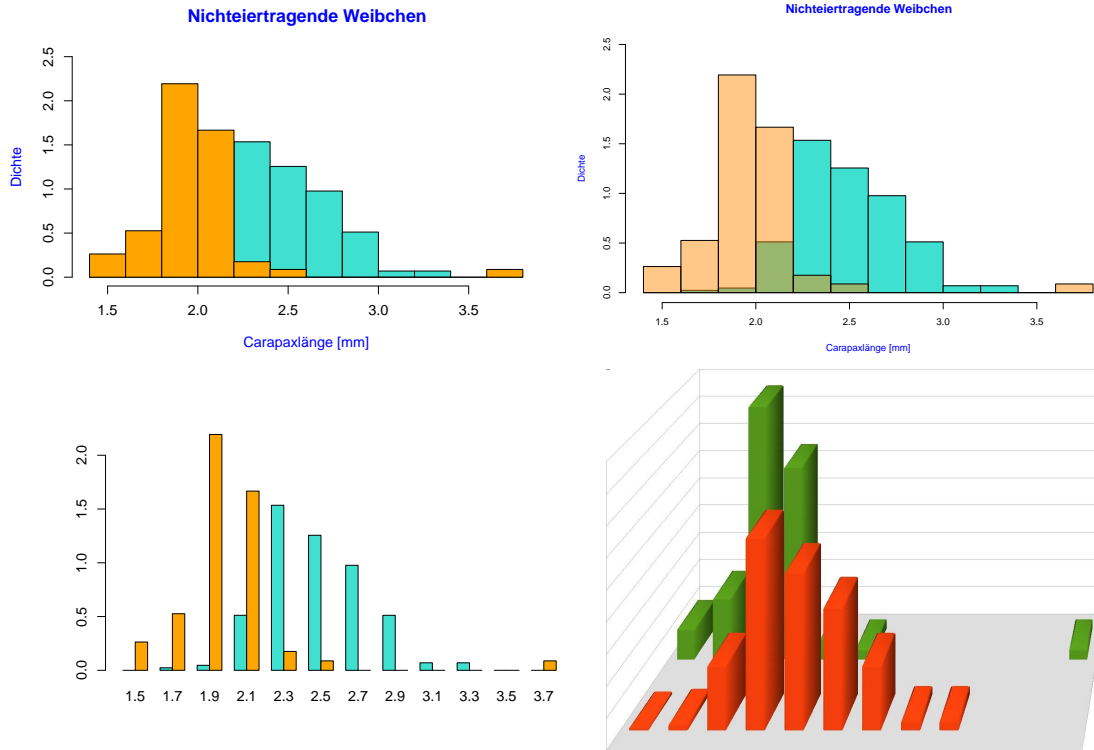


Nichteiertragende Weibchen am 6. Sept. '88, n=215



Die beiden Histogramme sind jetzt vergleichbar  
(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:

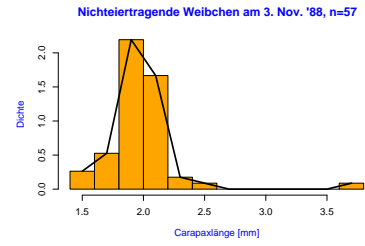
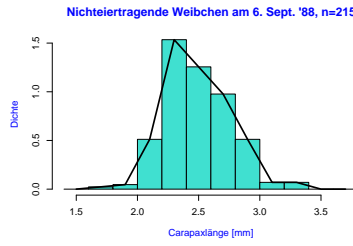
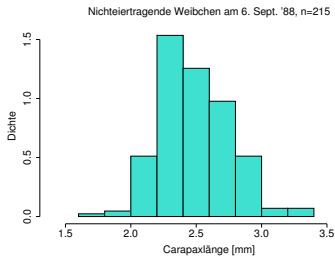
Bevorzugen Sie einfache und klare 2D-Darstellungen.

**Problem**

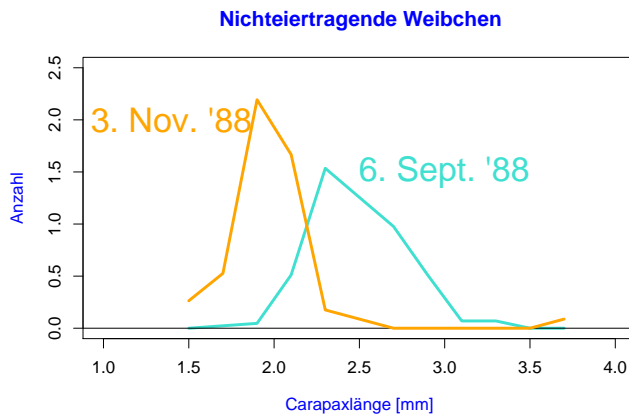
Histogramme kann man nicht ohne weiteres  
in demselben Graphen  
darstellen,  
weil sie einander  
überdecken würden.



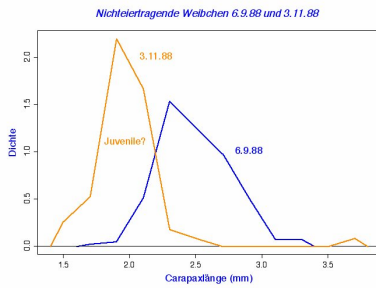
## Einfache und klare Lösung: Dichtepolygone



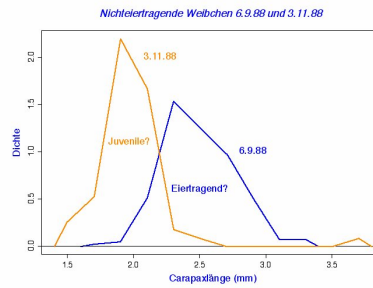
## Zwei und mehr Dichtepolygone in einem Plot



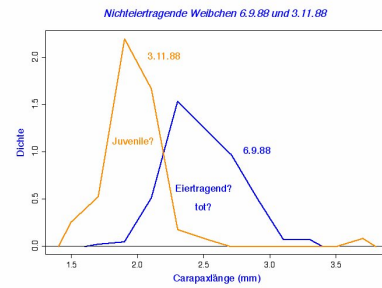
Biologische Interpretation der Verschiebung?



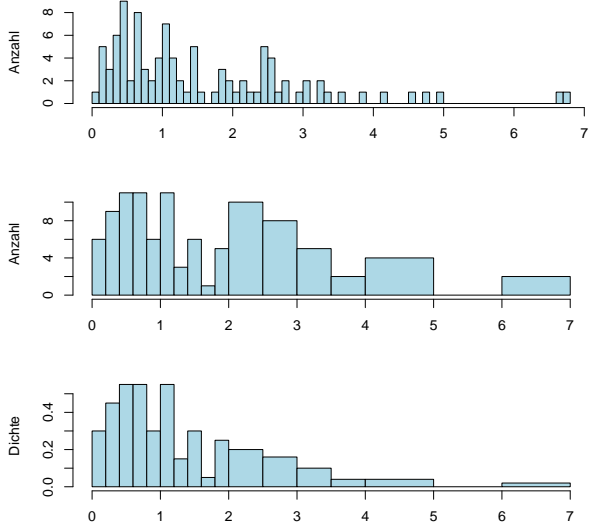
||



||

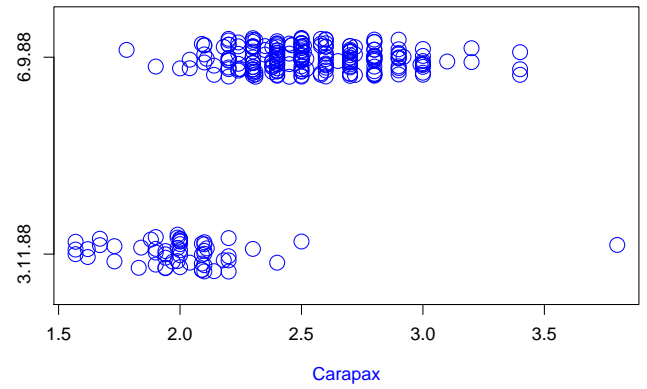
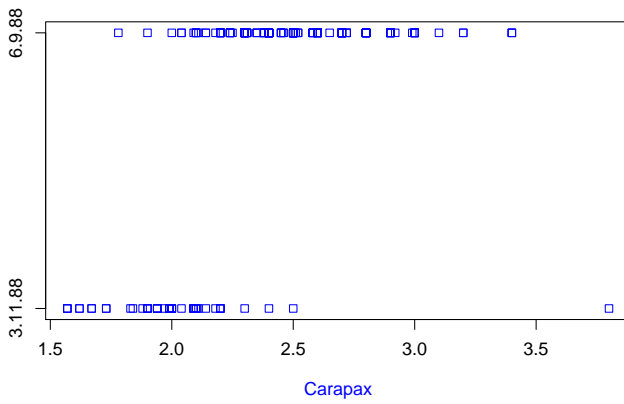


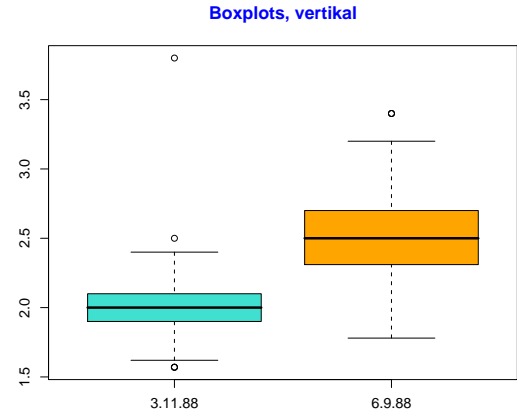
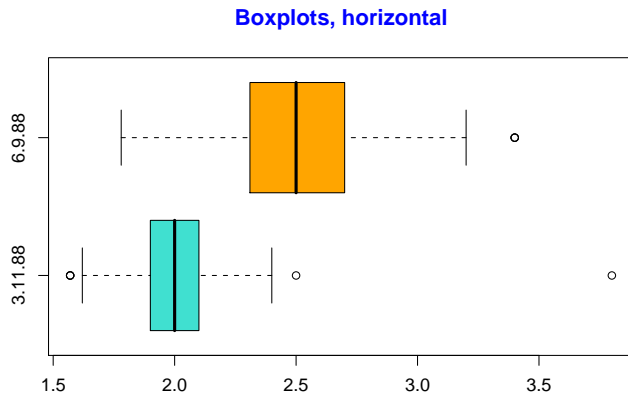
## Anzahl vs. Dichte



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

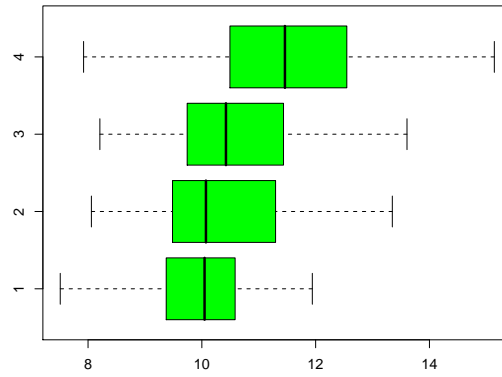
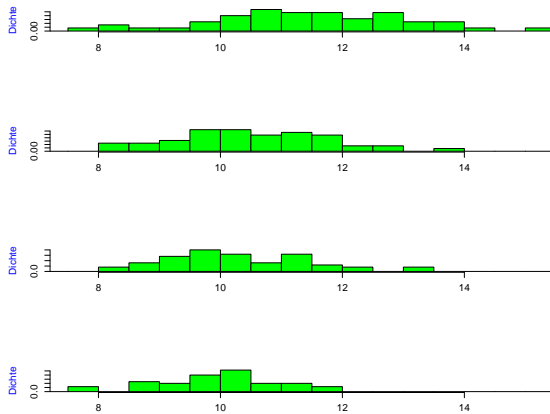


Baum Baum Baum Baum Baum Baum Baum

Baum Baum Baum Baum Baum Baum Baum

Wald?

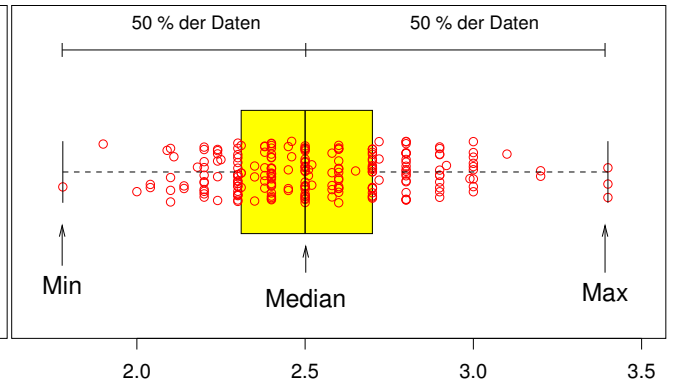
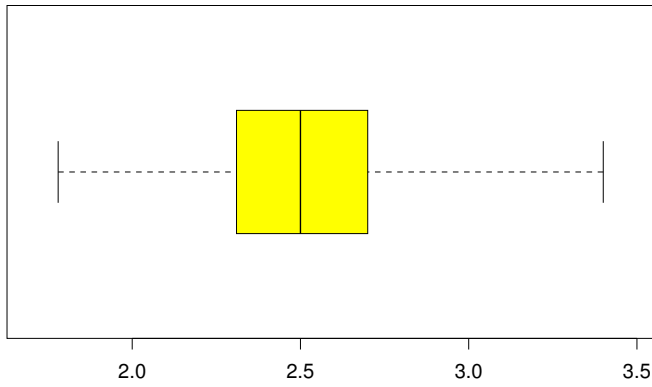
Beispiel:  
Vergleich von mehreren Gruppen



Der Boxplot

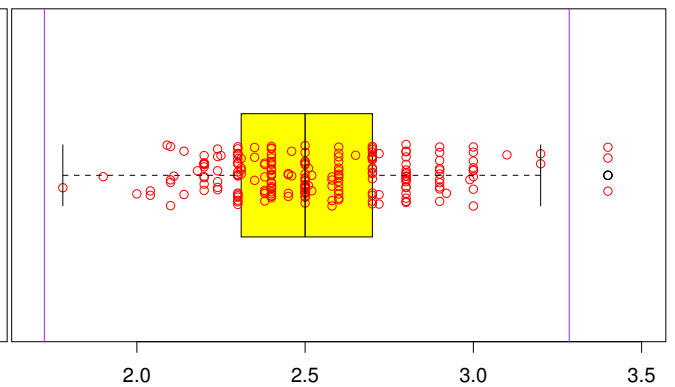
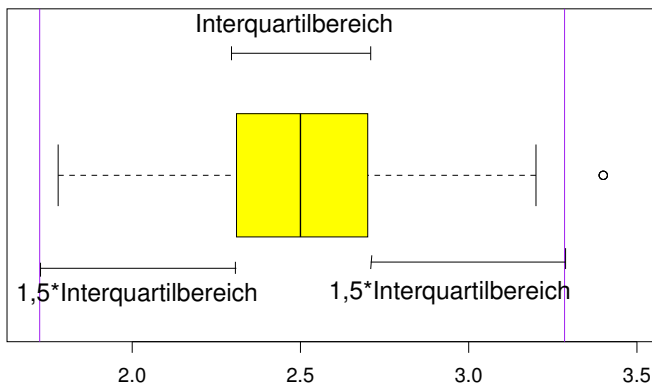
Boxplot, einfache Ausführung

Boxplot, einfache Ausführung



Carapaxlänge [mm]  
Boxplot, Standardausführung

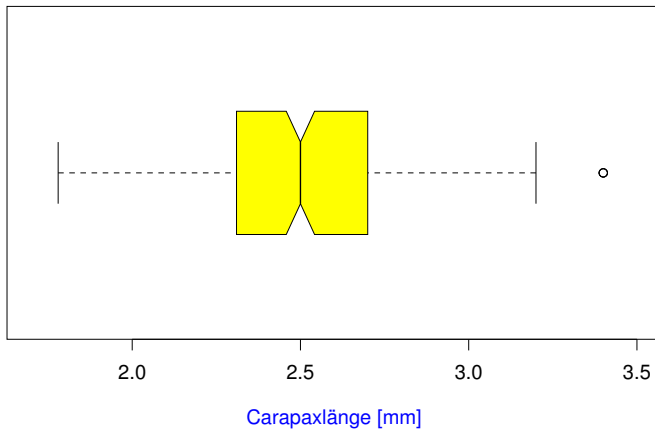
Carapaxlänge [mm]  
Boxplot, Standardausführung



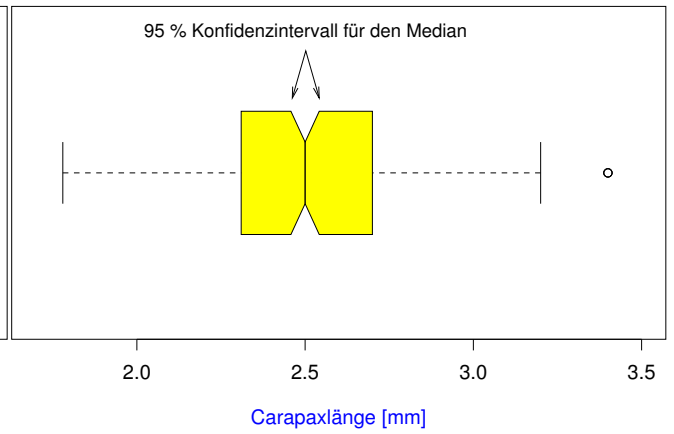
Carapaxlänge [mm]

Carapaxlänge [mm]

Boxplot, Profiausstattung



Boxplot, Profiausstattung



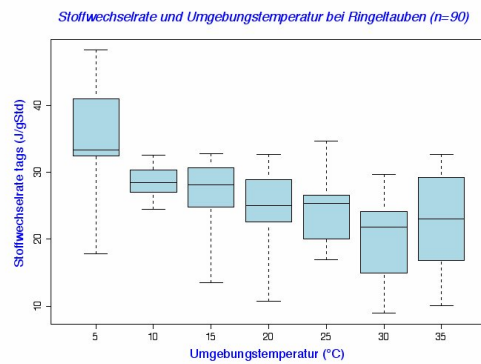
### 3.4 Beispiel: Ringeltaube

Beispiel:

Die Ringeltaube  
*Palumbus palumbus*

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

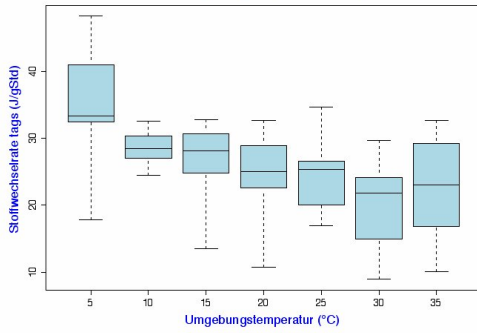
Daten aus dem AK Stoffwechselphysiologie  
Prof. Prinzinger Universität Frankfurt



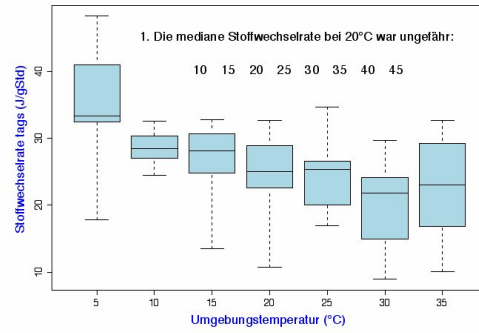
Klar: Stoffwechselrate *höher* bei *tiefen* Temperaturen

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

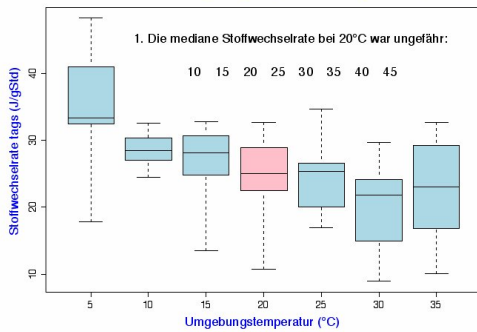
Stoffwechselrate und Umgebungstemperatur bei Ringellauben (n=90)



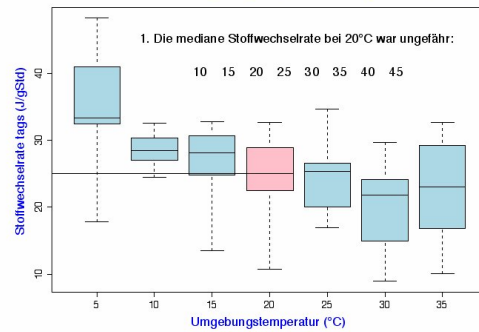
Stoffwechselrate und Umgebungstemperatur bei Ringellauben (n=90)



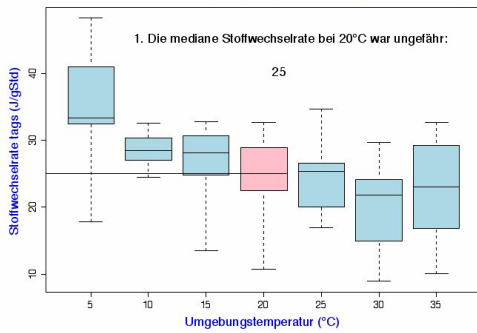
Stoffwechselrate und Umgebungstemperatur bei Ringellauben (n=90)



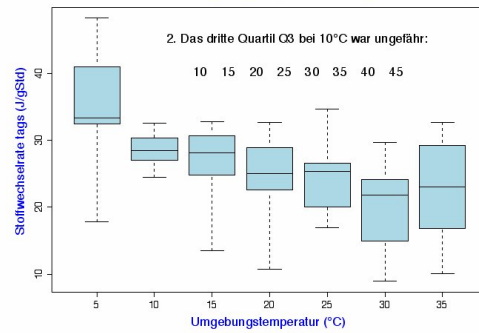
Stoffwechselrate und Umgebungstemperatur bei Ringellauben (n=90)



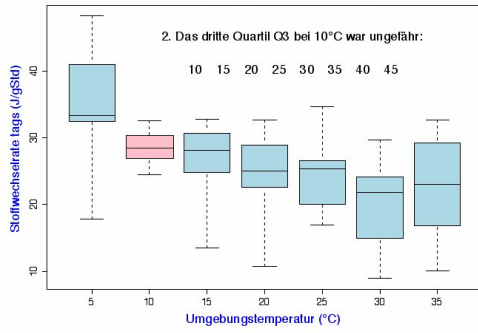
Stoffwechselrate und Umgebungstemperatur bei Ringellauben (n=90)



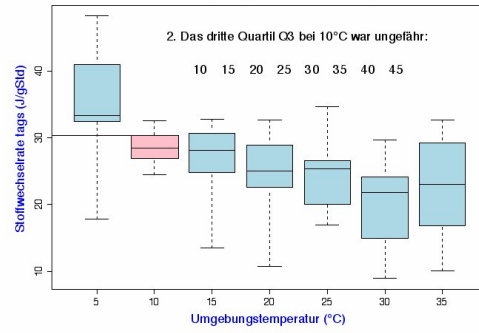
Stoffwechselrate und Umgebungstemperatur bei Ringellauben (n=90)



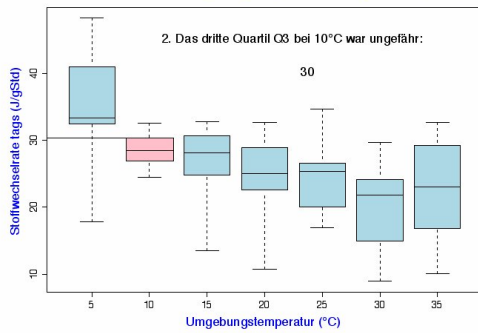
Stoffwechselrate und Umgebungstemperatur bei Ringellauben (n=90)



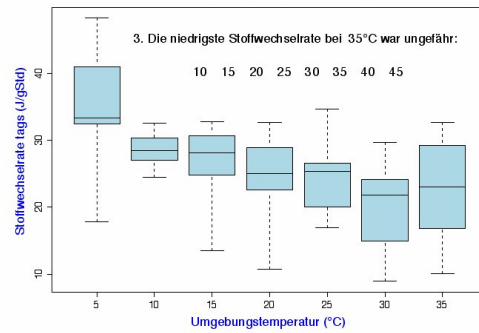
Stoffwechselrate und Umgebungstemperatur bei Ringellauben (n=90)



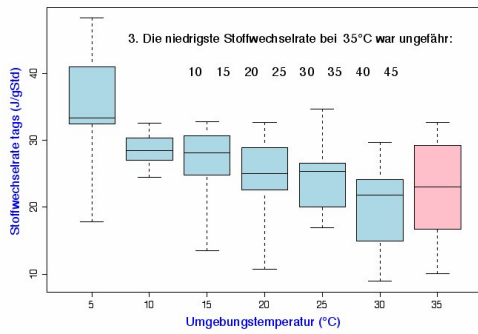
Stoffwechselrate und Umgebungstemperatur bei Ringellauben (n=90)



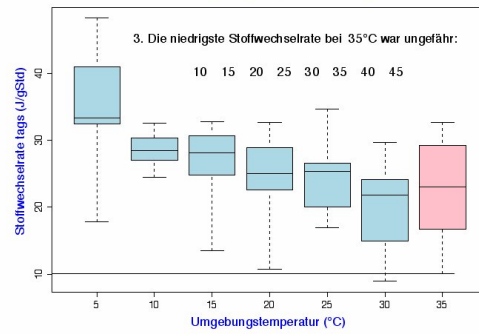
Stoffwechselrate und Umgebungstemperatur bei Ringellauben (n=90)



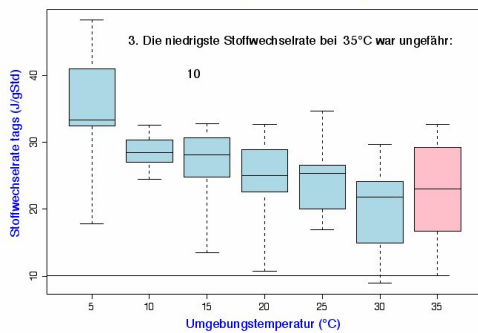
Stoffwechselrate und Umgebungstemperatur bei Ringellauben (n=90)



Stoffwechselrate und Umgebungstemperatur bei Ringellauben (n=90)



Stoffwechselrate und Umgebungstemperatur bei Ringellauben (n=90)

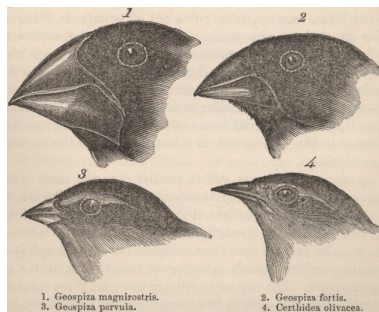


### 3.5 Beispiel: Darwin-Finken

Charles Robert Darwin (1809-1882)



#### Darwin-Finken



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

#### Darwins Finken-Sammlung

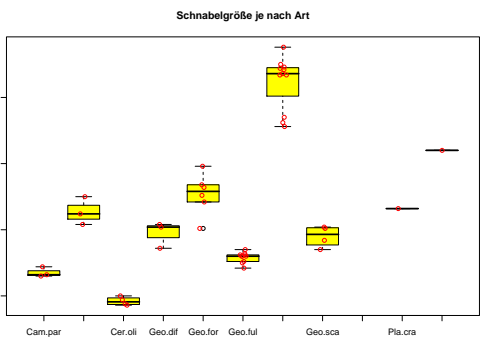
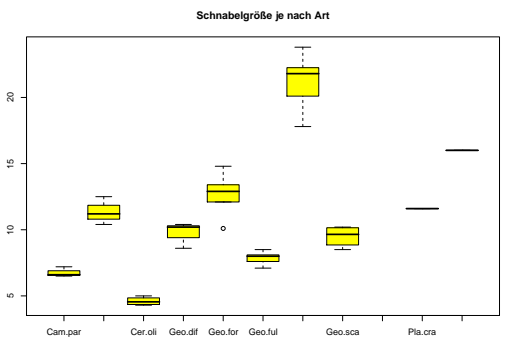
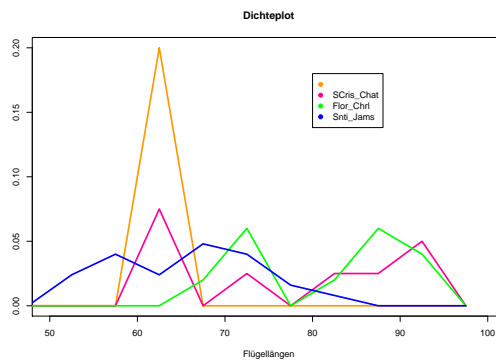
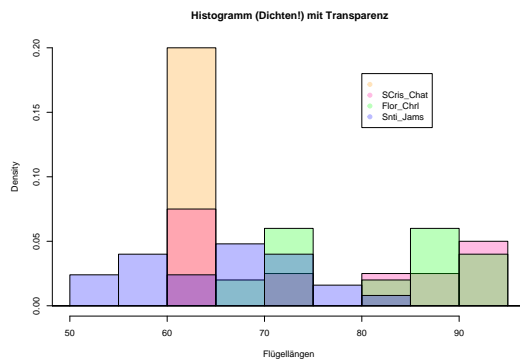
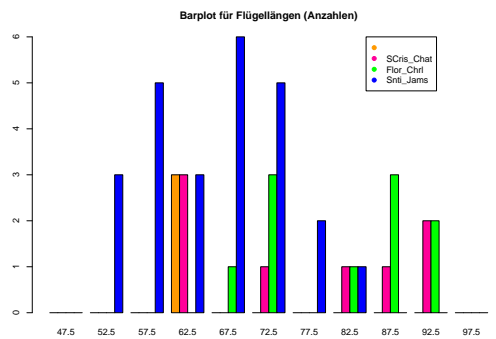
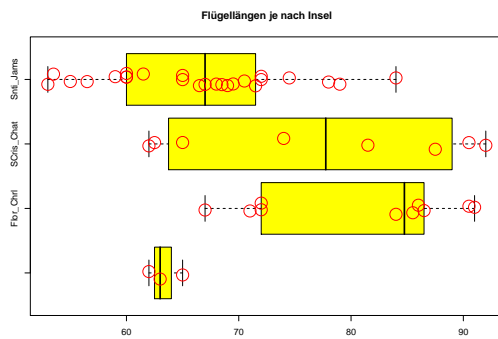
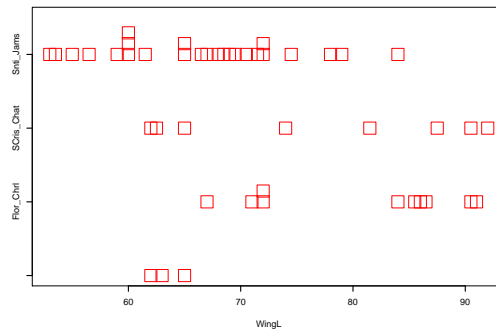
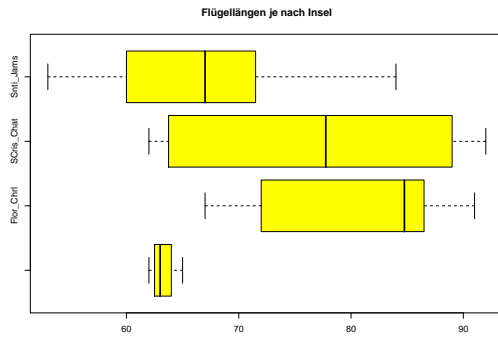
#### Literatur

[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>

#### Flügelängen der Darwin-Finken





**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

## 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*  $\bar{x}$

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[1ex] = [1ex]

Summe der Messwerte  
Anzahl der Messwerte

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

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:

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

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

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

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

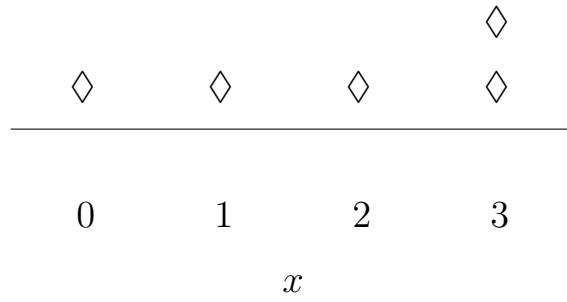
$$\bar{x} = 1,8$$

Geometrische Bedeutung des Mittelwerts:

Der Schwerpunkt

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

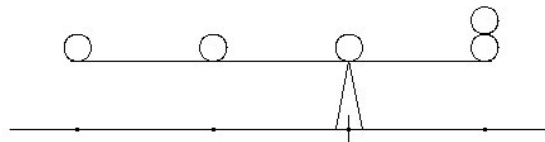
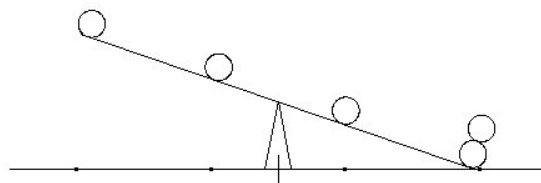
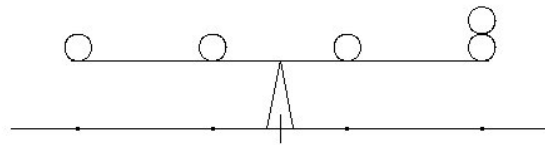
Wo muß der Drehpunkt sein, damit die Waage im Gleichgewicht ist?

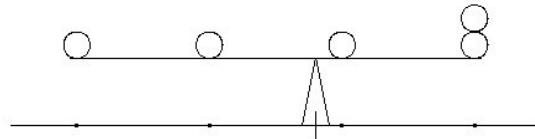
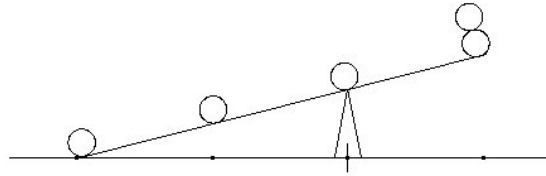


$m = 1,5 ?$

$m = 2 ?$

$m = 1,8 ?$





zu klein

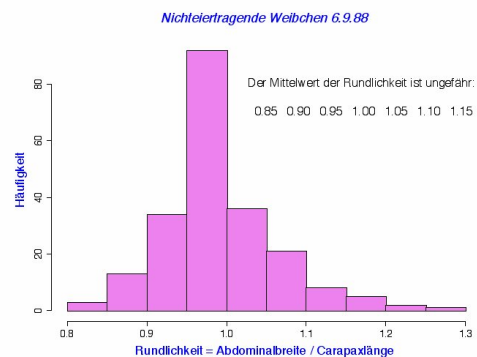
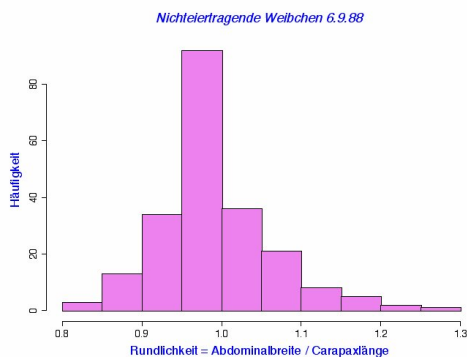
zu groß

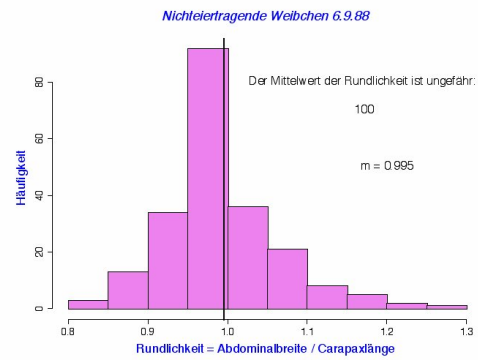
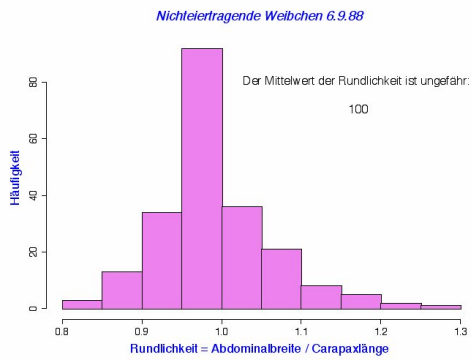
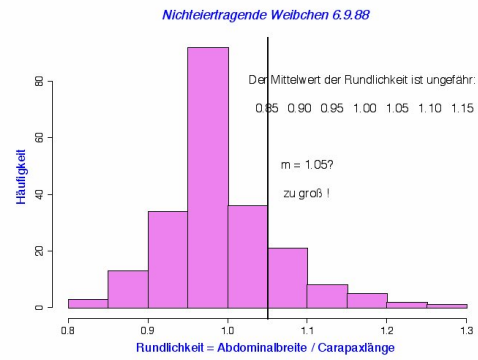
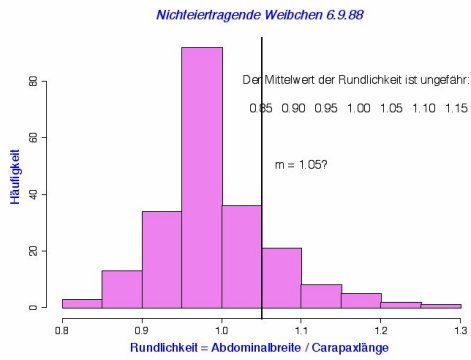
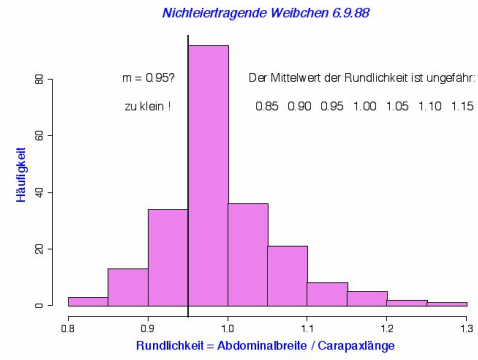
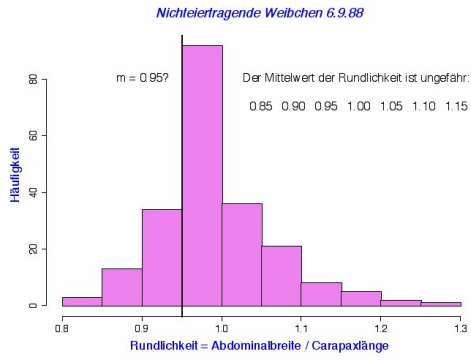
richtig

Beispiel: *Galathea intermedia*

„Rundlichkeit“ :=  $\text{Abdominalbreite} / \text{Carapaxlänge}$

Vermutung: Rundlichkeit nimmt bei Geschlechtsreife zu

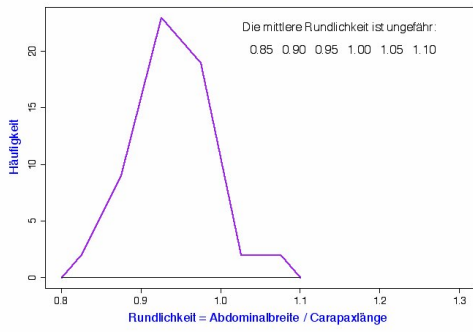




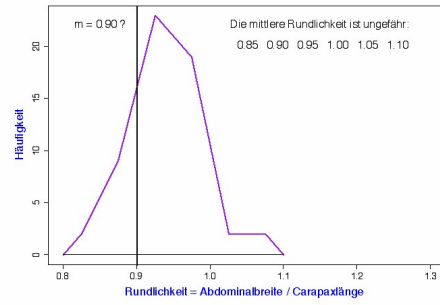
Beispiel:

3.11.88

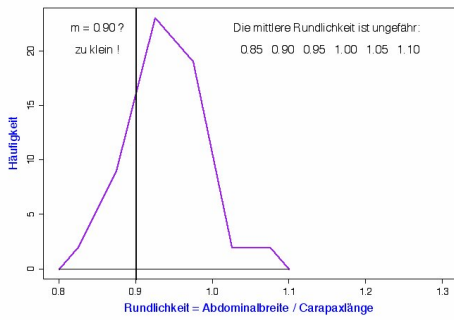
Nichteiertragende Weibchen 3.11.88



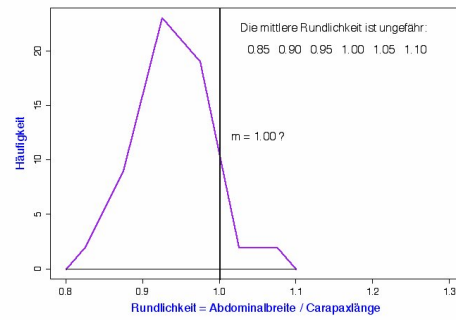
Nichteiertragende Weibchen 3.11.88



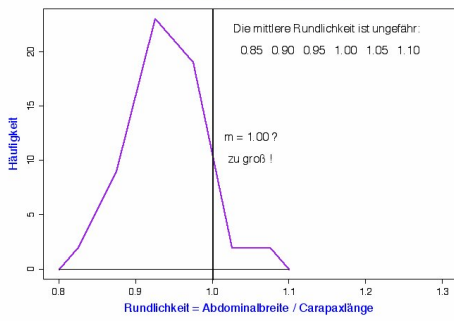
Nichteiertragende Weibchen 3.11.88



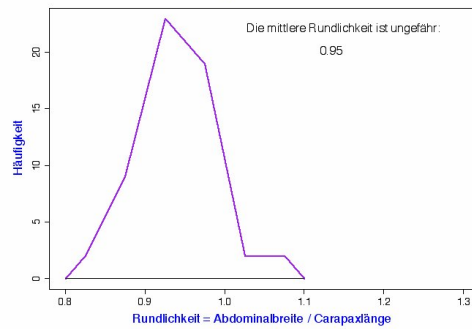
Nichteiertragende Weibchen 3.11.88



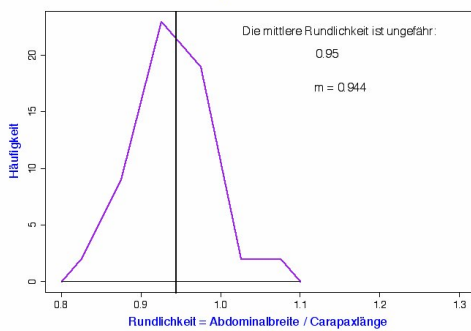
Nichteiertragende Weibchen 3.11.88



Nichteiertragende Weibchen 3.11.88



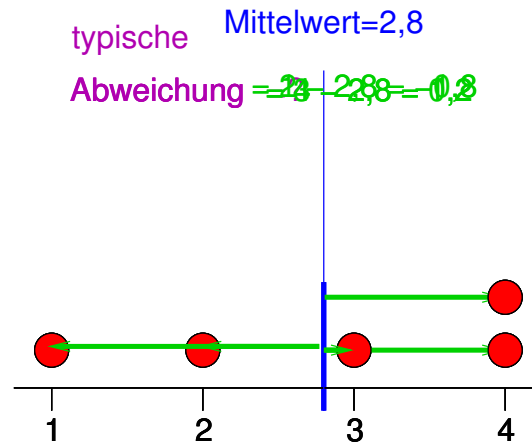
Nichteiertragende Weibchen 3.11.88





## Die Standardabweichung

Wie weit weicht eine typische Beobachtung vom Mittelwert ab ?



Die *Standardabweichung*  $\sigma$  (“sigma”) ist ein etwas komisches gewichtetes Mittel der Abweichungsbeträge und zwar

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

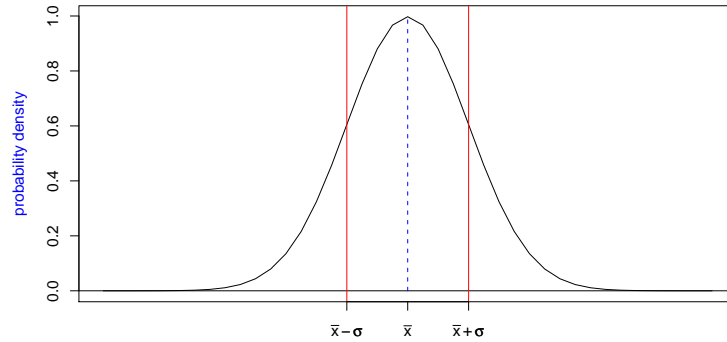
Die *Standardabweichung* von  $x_1, x_2, \dots, x_n$  als Formel:

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

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

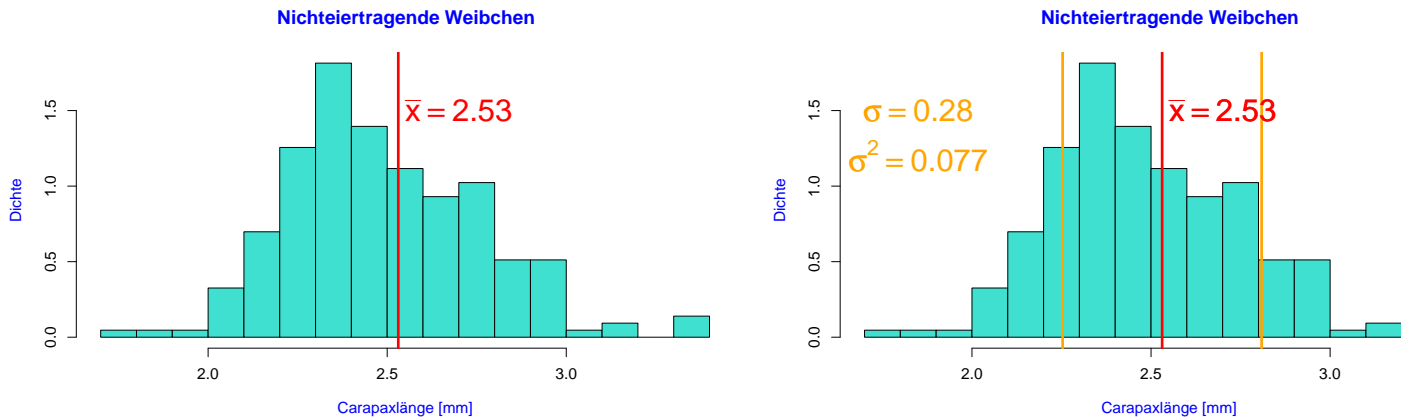
**Faustregel für die Standardabweichung**

Bei ungefähr glockenförmigen (also eingipfligen und symmetrischen) Verteilungen liegen ca. 2/3 der Ver-



teilung 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 nichteiertragender Weibchen vom 6.9.88

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

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

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

### 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 meistens das korrigierte  $s$  gemeint.

Beispiel Die Daten  $\bar{x} = ?$   $\bar{x} = 10/5 = 2$   
*Summe*

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

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