Statistics for EES

General Introduction and Descriptive Statistics

Dirk Metzler

April 11, 2020

Contents

Contents

1 Intro: What is Statistics?

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

Andrejs Dunkels

What is Statistics?

Nature is full of Variability

How to make sense of variable data?

Use mathematical theory of randomness:[0.5ex] *Probability.*

Statistics

=

Data Analysis

based on

Probabilistic Models

Some of the aims of this course

- Understand the priciples underlying statistics and probability
- Understand widely used statistical methods
- Learn to apply these methods to data (with R)
- Understand under which conditions these methods work, and under which conditions they do not and why
- Learn when to choose which method and when to consult an expert
- Be able to read an judge scientific publications in which non-standard statistical methods are applied and explained
- Get a feel of randomness

How to study the content of the lecture

For the case that you are overwhelmed by the contents of this course, and if you don't have a good strategy to study, here is my recommendation:

1. Try to explain the items under "Some of the things you should be able to explain"
2. Discuss these explanations with your fellow students
3. Do this before the next lecture, such that you can ask questions if things don't become clear
4. Do the exercises in time and present your solutions
5. Study all the rest from the handout, your notes during the lecture, and in books

ECTS and work load per week

3 ECTS correspond to $\frac{3\times 30}{14} \approx 6.43$ hours of work per week, e.g.

- 2.4 hours spent in lectures and exercise sessions
- 1.5 hours of revising the contents of the lecture
- 2.5 hours of solving exercise problems (including data analyses and theoretical problems)

What will the exam be like

You can bring:

- pocket calculator
- formula sheet, hand-written by yourself

What you need to answer the questions:

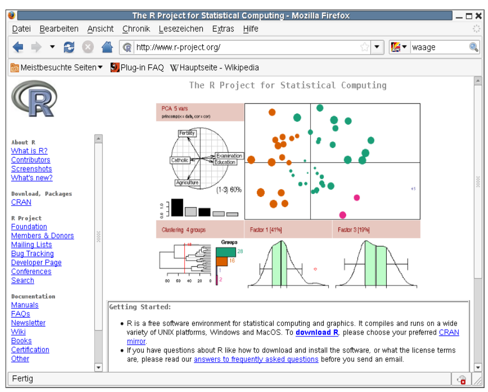
- understanding concepts
- be able to apply concepts
- do calculations
- think during the exam
- (not just reproduce facts)
- have done the exercise sheets and discussed the solutions!

Descriptive Statistics

Descriptive Statistics is

the first look at the data.

Statistics Software R



http://www.r-project.org

2 Data Visualization

Data Example

Data from a biology diploma thesis, 2001, Forschungsinstitut Senckenberg, Frankfurt am Main

Crustacea section

Advisor: Prof. Dr. Michael Türkay

Charybdis acutidens TÜRKAY 1985

The Squat Lobster

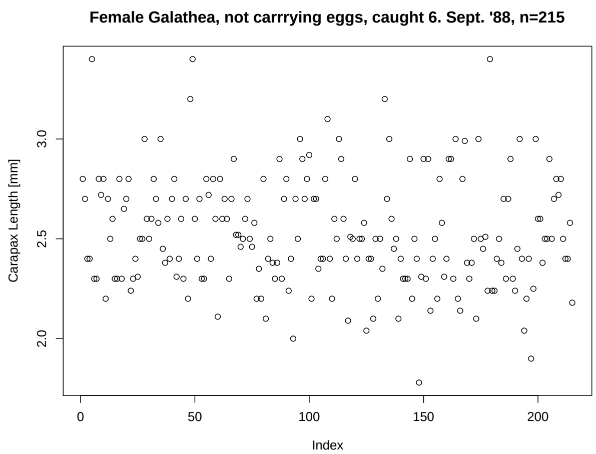
Galathea intermedia

Squat Lobsters, caught 6. Sept 1988

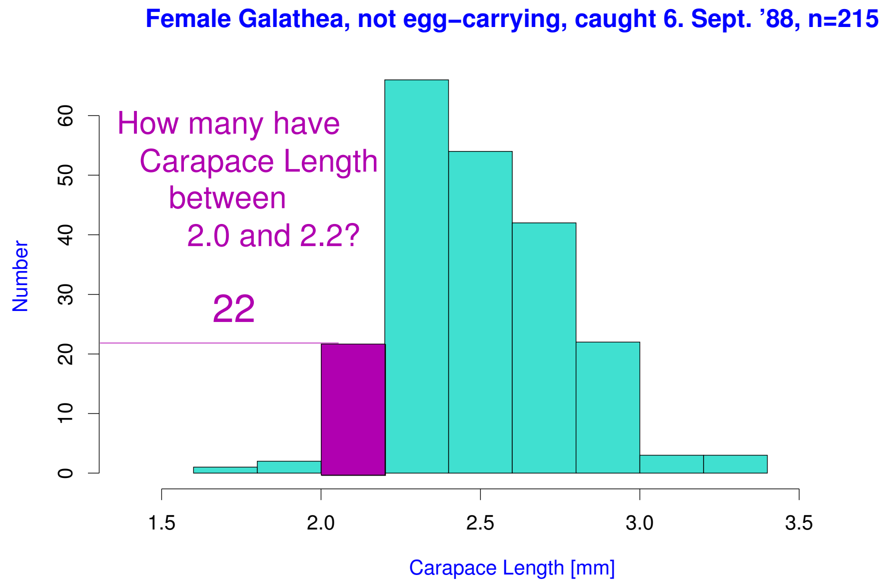
Helgoländer Tiefe Rinne, North Sea

Carpace Lengths (mm): Females, not egg-carrying ($n = 215$)

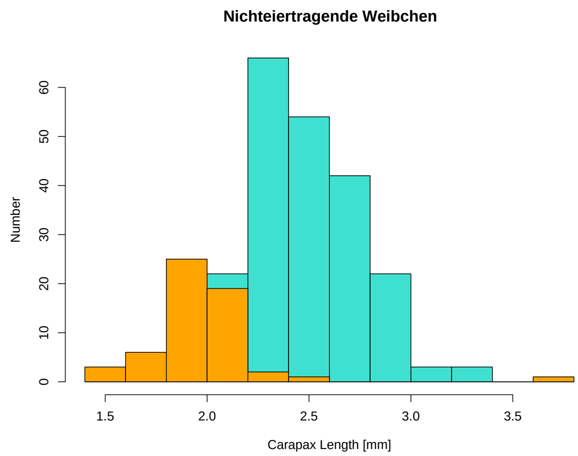
2.9	3.0	2.9	2.5	2.7	2.9	2.9	3.0
3.0	2.9	3.4	2.8	2.9	2.8	2.8	2.4
2.8	2.5	2.7	3.0	2.9	3.2	3.1	3.0
2.7	2.5	3.0	2.8	2.8	2.8	2.7	3.0
2.6	3.0	2.9	2.8	2.9	2.9	2.3	2.7
2.6	2.7	2.5	.	.	.	.	.



2.1 Histograms und Density Polygons



Comparing the two Distributions

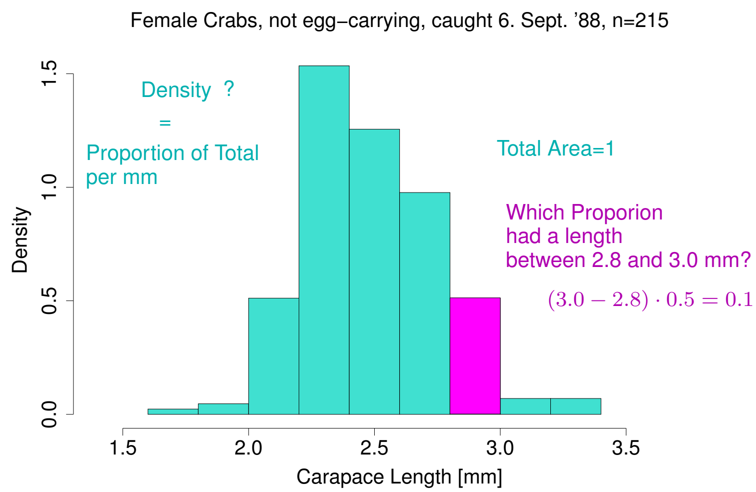


Problem: different sample sizes

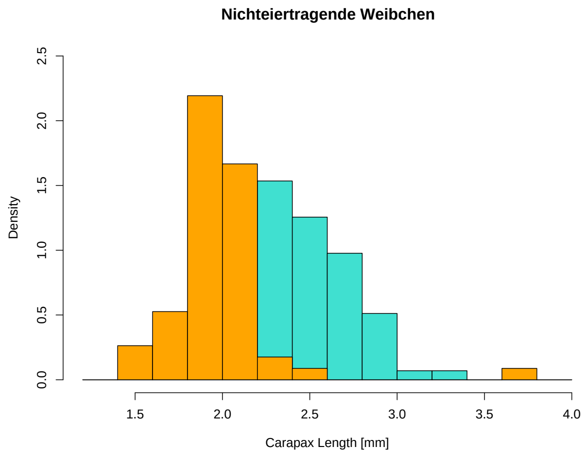
6.9.1988 : $n = 215$

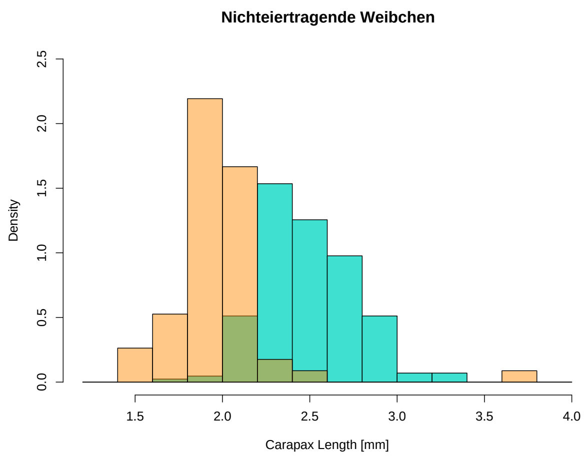
3.11.1988 : $n = 57$

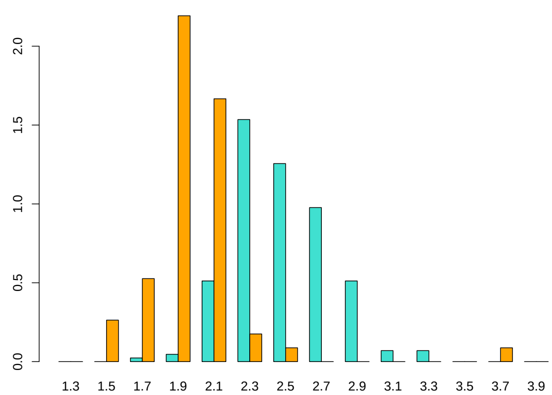
Idea: scale y-axis such that each distribution has total area 1.

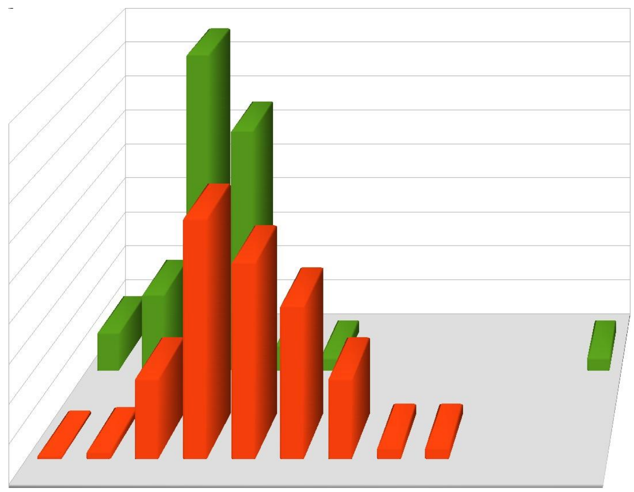

Female Crabs, not egg−carrying, caught 6. Sept. '88, n=215
Density ?
=
Proportion of Total
per mm
Total Area=1
Which Proporion
had a length
between 2.8 and 3.0 mm?
$(3.0 - 2.8) \cdot 0.5 = 0.1$
Density
0.0
0.5
1.0
1.5
1.5
2.0
2.5
3.0
3.5
Carapace Length [mm]

How to compare the two distributions?









My Advice

If you are a commercial artist:

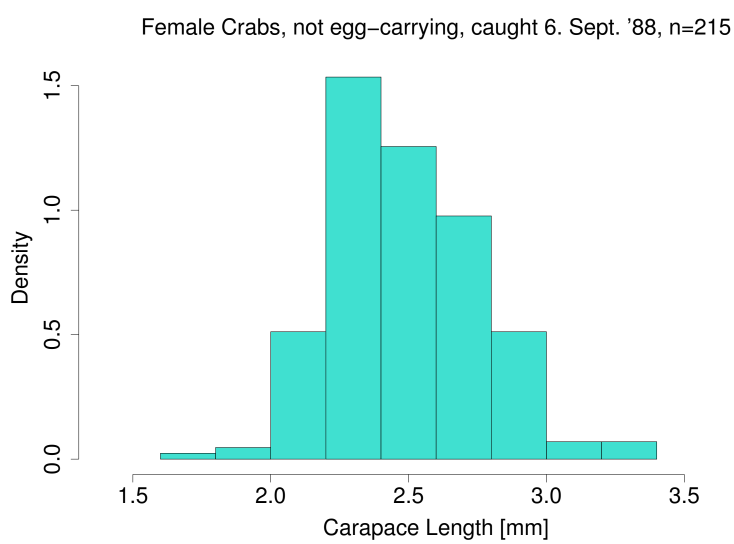
Impress everybody with cool 3D graphics!

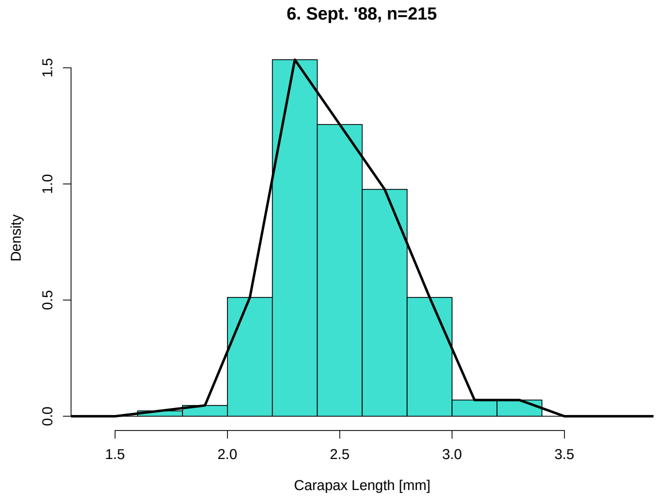
If you are a scientist:

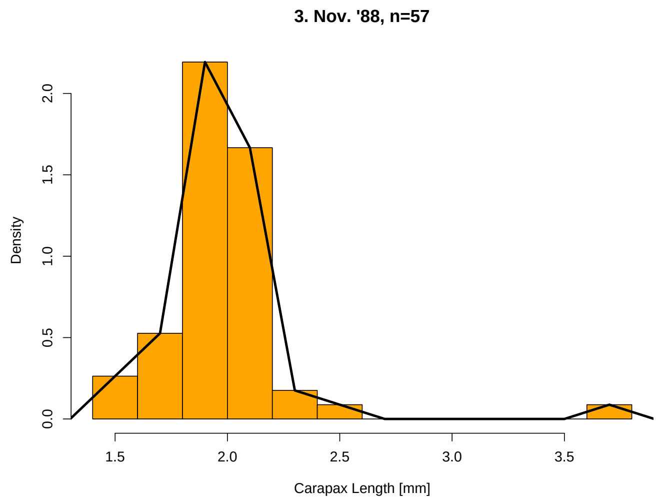
Visualize your data in clear and simple 2D plots.

(As long as you print on 2D paper and project your slides on 2D screens)

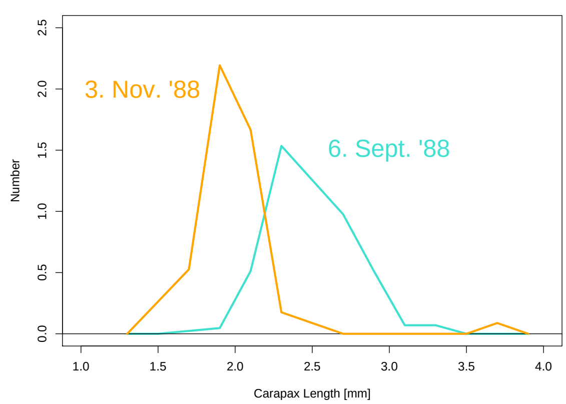
Simple and Clear: Density Polygons







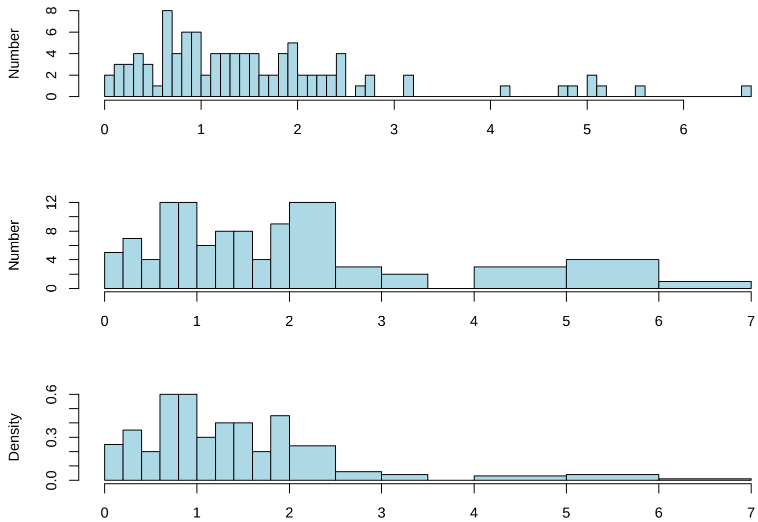
Convenient to show two or more Density Polygons in one plot



Biological Interpretation: What may be the reason for this shift?

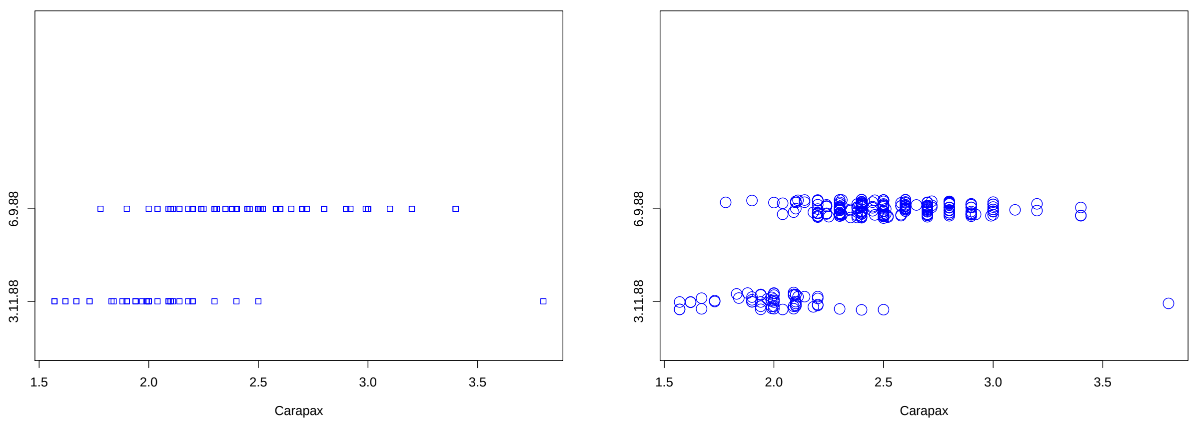
2.1.1 Histograms: Densities or Numbers?

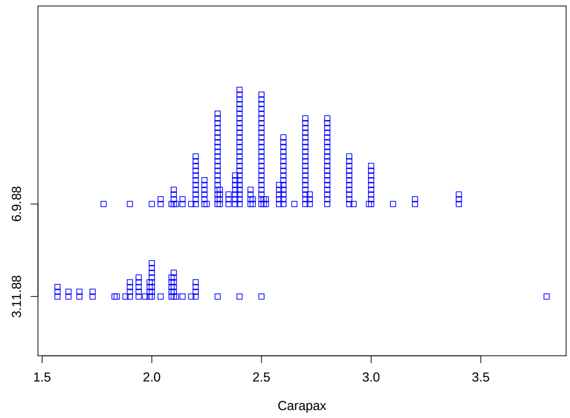
Number vs. Density

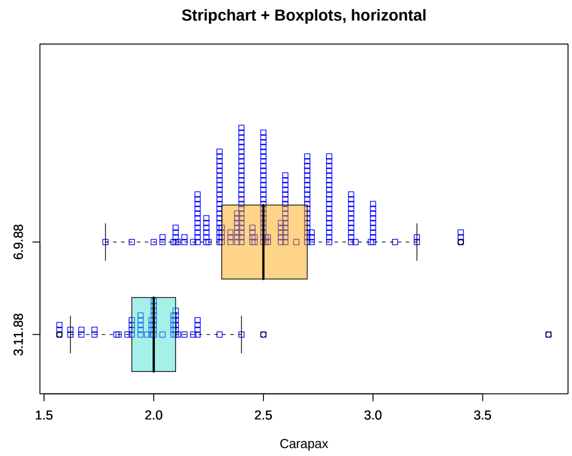


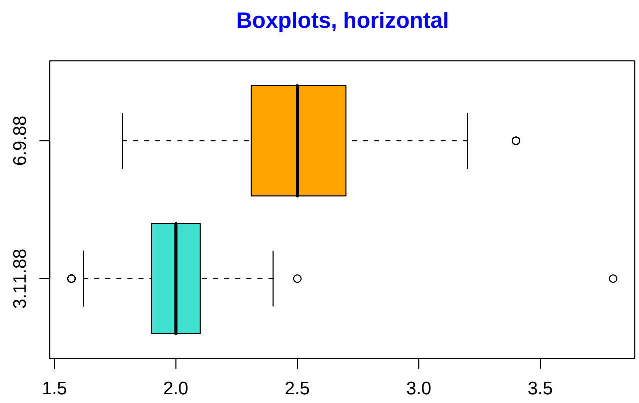
Histograms with unequal intervals should show densities, not numbers!

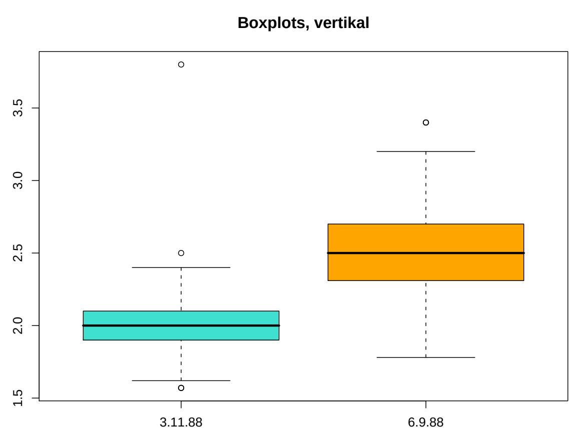
2.2 Stripcharts and Boxplots









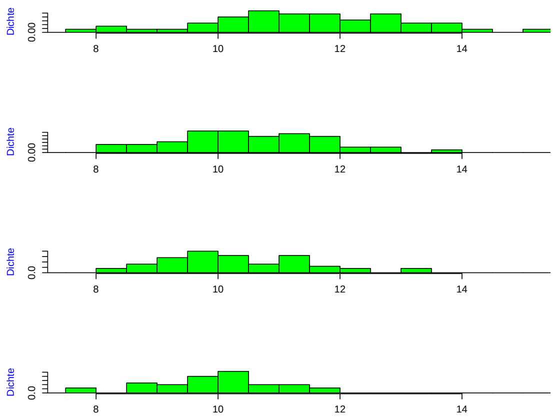


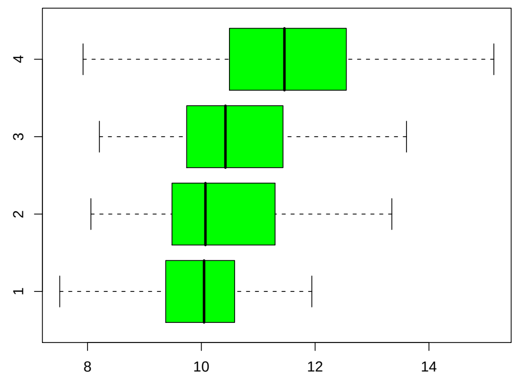
Simplify to understand

Histograms and density polygons
allow a comprehensive view on the data.

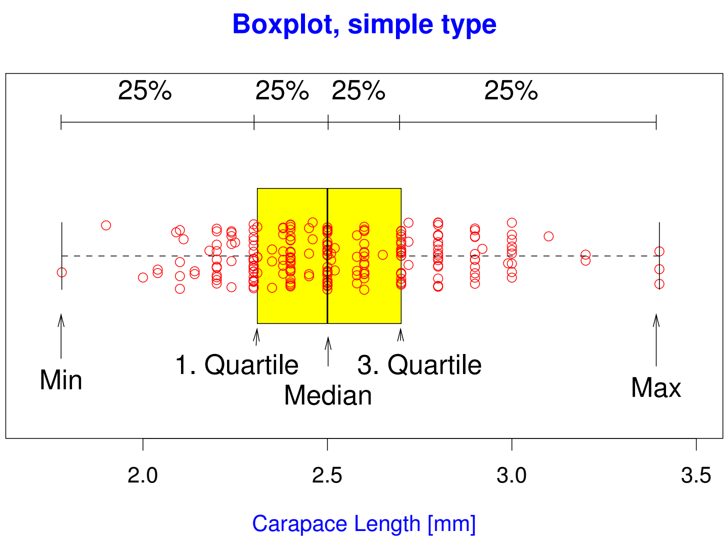
Sometimes too comprehensive.

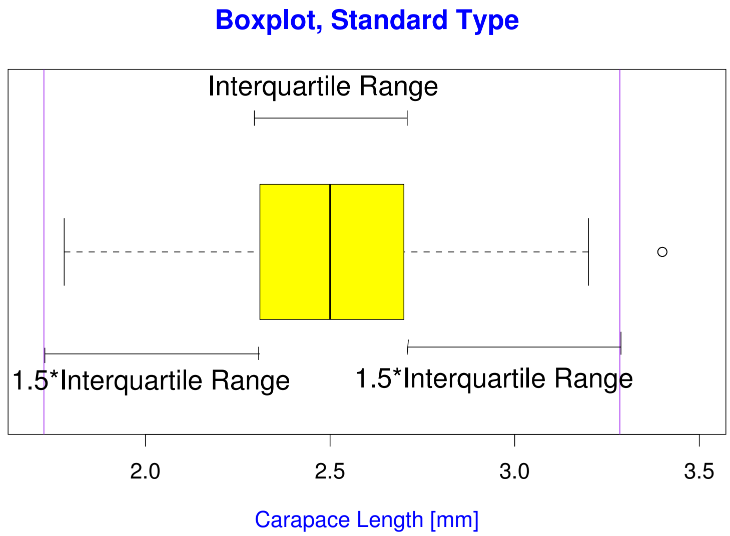
Comparison of four groups

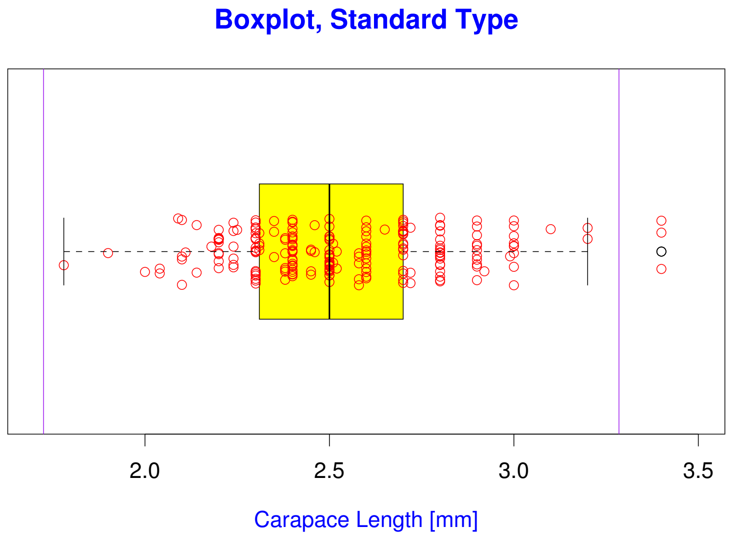


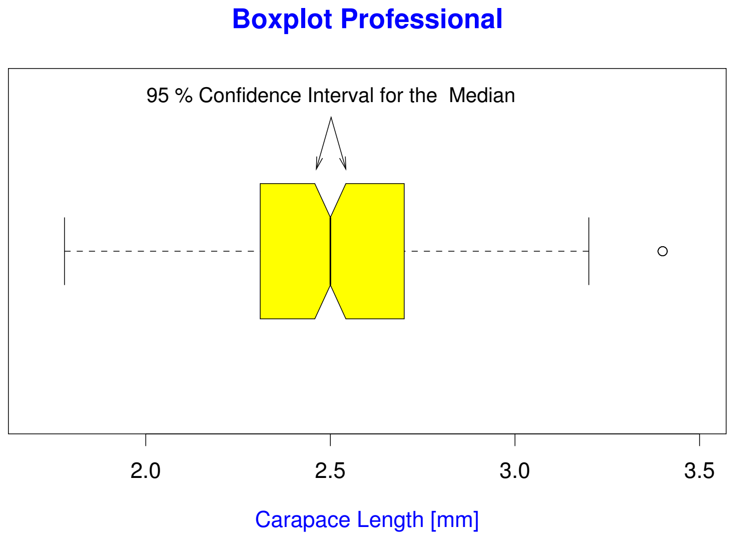


The Boxplot




Boxplot, Standard Type
Interquartile Range
1.5*Interquartile Range
1.5*Interquartile Range
2.0
2.5
3.0
3.5
Carapace Length [mm]


Boxplot, Standard Type
2.0
2.5
3.0
3.5
Carapace Length [mm]


Boxplot Professional
95 % Confidence Interval for the Median
2.0
2.5
3.0
3.5
Carapace Length [mm]

Example: Darwin Finches

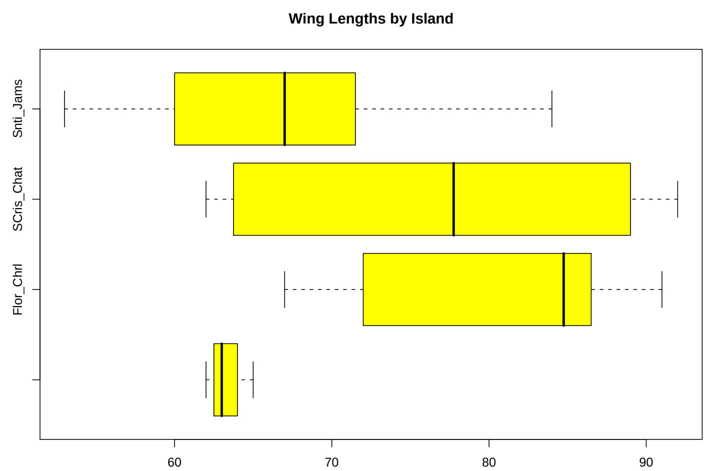
Darwin's collection of Finches

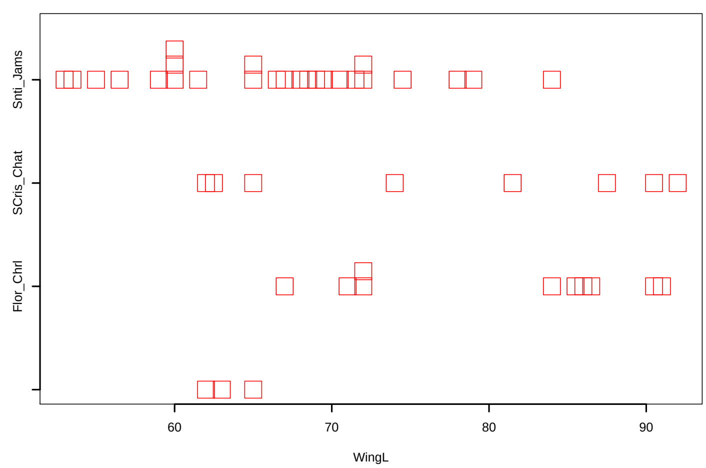
References

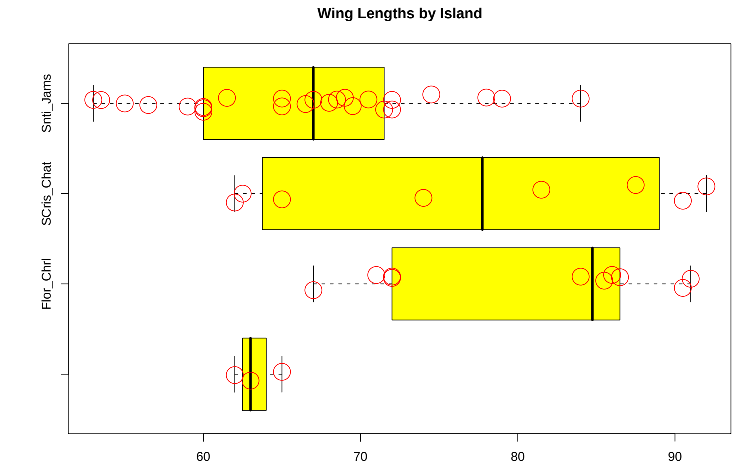
[1] Sulloway, F.J. (1982) The Beagle collections of Darwin's Finches (Geospizinae). *Bulletin of the British Museum (Natural History), Zoology series* **43**: 49-94.

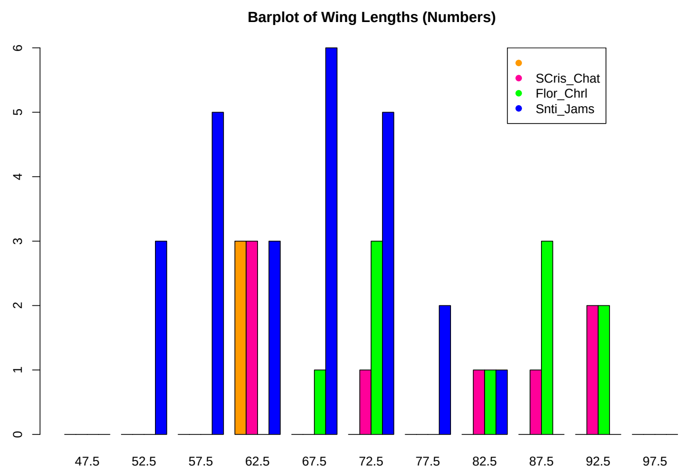
[2] http://datadryad.org/repo/handle/10255/dryad.154

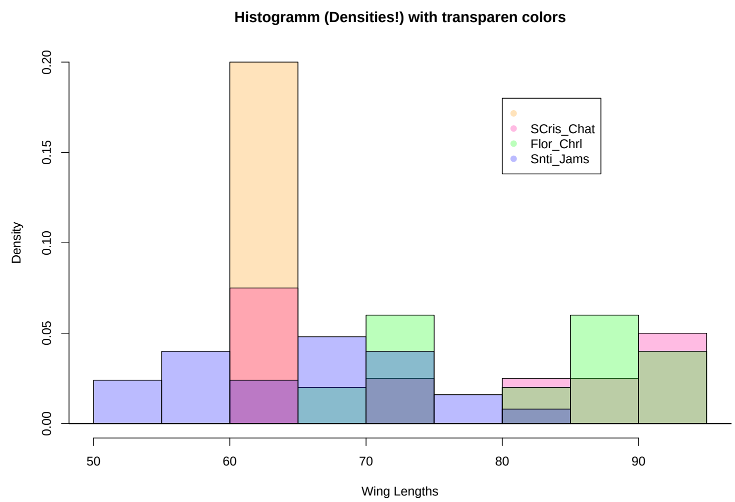
Wing Sizes of Darwin's Finches

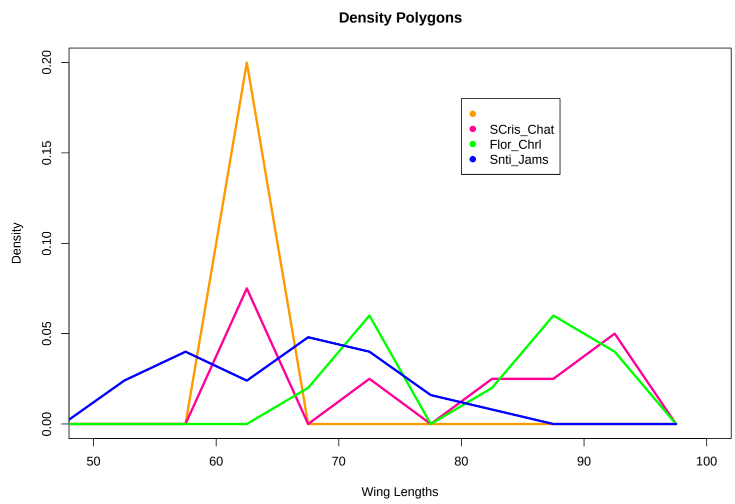




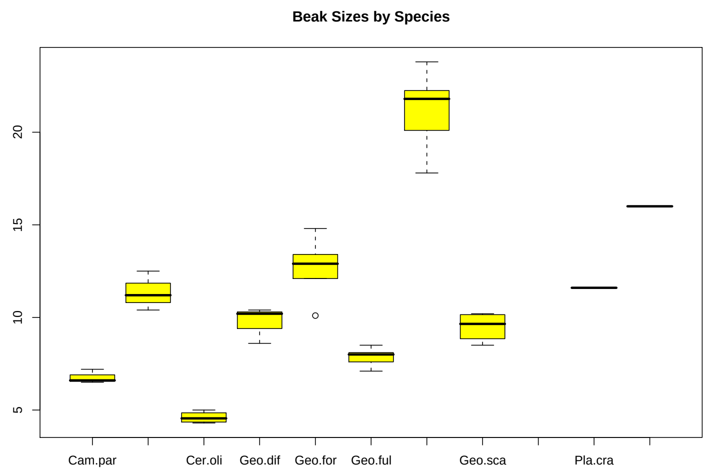

Wing Lengths by Island
Snti_Jams
SCris_Chat
Flor_Chrl
60
70
80
90

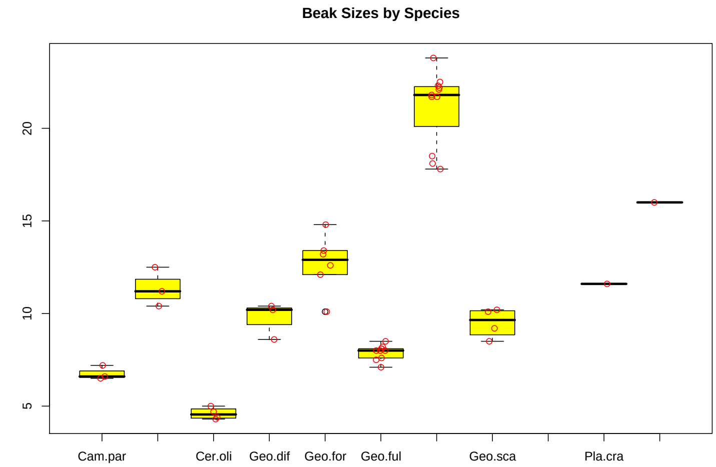

Barplot of Wing Lengths (Numbers)
SCris_Chat
Flor_Chrl
Snti_Jams
0
1
2
3
4
5
6
47.5
52.5
57.5
62.5
67.5
72.5
77.5
82.5
87.5
92.5
97.5


Histogramm (Densities!) with transparen colors
SCris_Chat
Flor_Chrl
Snti_Jams
Density
0.00
0.05
0.10
0.15
0.20
50
60
70
80
90
Wing Lengths



Beak Sizes of Darwin's Finches



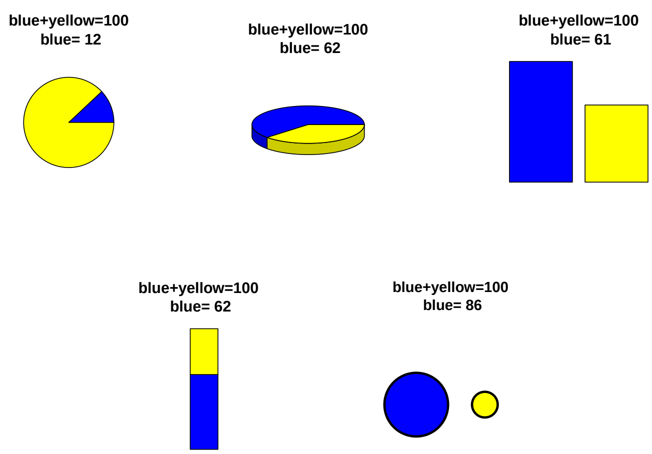


2.3 Conclusions

Conclusions

- Histograms give detailed information.
- Density Polygons allow multiple comparisons.
- Boxplots can simplify large datasets.
- Stripcharts more appropriate for small datasets.
- Sophisticated graphics with 3D or semi-transperent colors do not always improve clarity.

2.4 Pie charts or bar charts? An experiment



2.5 Example: blue area challenge results from 2018

Reading the data

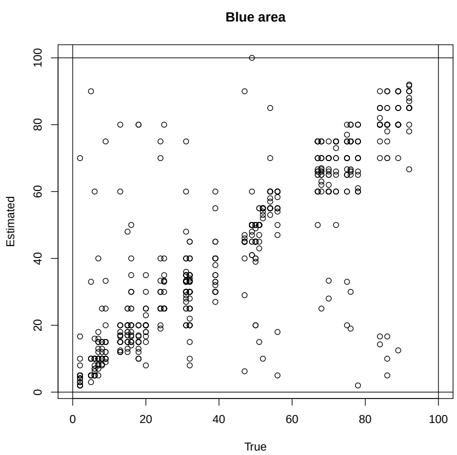
```
est <- read.csv("DataAndR/bluearea_estimates_2018.csv")
str(est)

## 'data.frame': 880 obs. of  5 variables:
##  $ Figure   : int  1 2 3 4 5 6 7 8 9 10 ...
##  $ type     : Factor w/ 5 levels "bar","bar.stack",..: 4 2 3 1 1 5 5 5 4 4 ...
##  $ estimated: num  50 65 100 25 75 10 30 NA 20 50 ...
##  $ student  : int  1 1 1 1 1 1 1 1 1 1 ...
##  $ true     : int  50 68 49 25 76 8 31 72 50 16 ...
```

```
head(est)

##   Figure      type estimated student true
## 1      1       pie        50       1   50
## 2      2 bar.stack        65       1   68
## 3      3      circ       100       1   49
## 4      4       bar        25       1   25
## 5      5       bar        75       1   76
## 6      6     pie3D        10       1    8
```

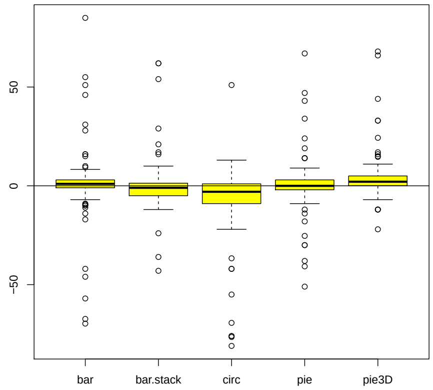
```
plot(est$true,est$estimate,xlab="True", ylab="Estimated",
     main="Blue area",xlim=c(0,100))
abline(h=c(0,100))
abline(v=c(0,100))
```



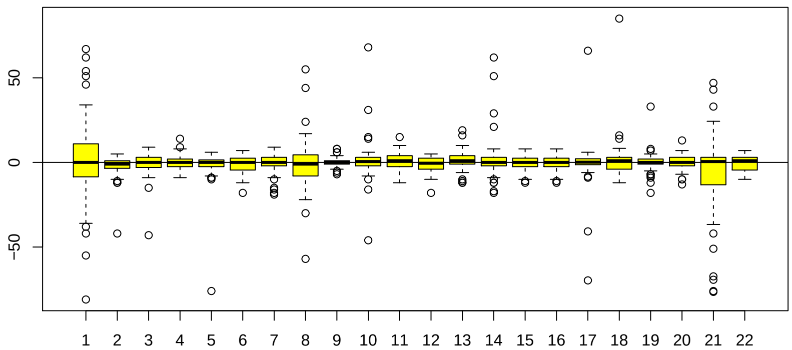

```
est$error <- est$estimate-est$true
str(est)

## 'data.frame': 880 obs. of  6 variables:
##  $ Figure   : int  1 2 3 4 5 6 7 8 9 10 ...
##  $ type     : Factor w/ 5 levels "bar","bar.stack",..: 4 2 3 1 1 5 5 5 4 4 ...
##  $ estimated: num  50 65 100 25 75 10 30 NA 20 50 ...
##  $ student  : int  1 1 1 1 1 1 1 1 1 1 ...
##  $ true     : int  50 68 49 25 76 8 31 72 50 16 ...
##  $ error    : num  0 -3 51 0 -1 2 -1 NA -30 34 ...
```

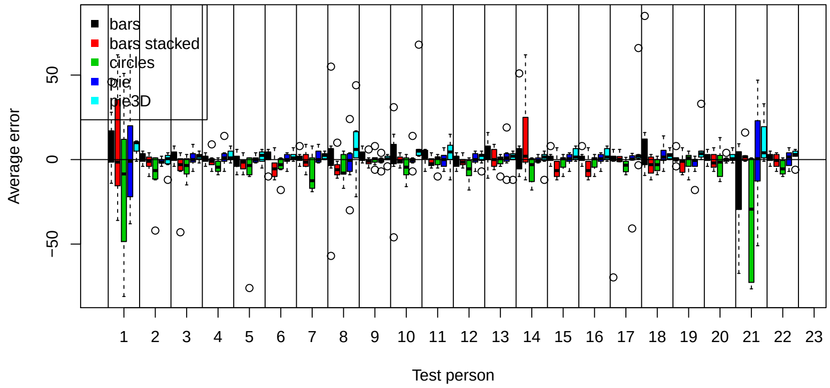
```
boxplot(error~type,est,col="yellow")
abline(h=0)
```




```
boxplot(error~student,est,col="yellow")
abline(h=0)
```




```
boxplot(error~type+student,est,col=rep(1:5,23),
        xaxt="none",xlab="Test person",ylab="Average error")
abline(v=0:24*5+0.5)
abline(h=0)
legend("topleft",pch=15,col=1:5,
       legend=c("bars","bars stacked","circles","pie","pie3D"))
axis(side=1,at=1:23*5-2,labels=1:23)
```



3 Summarizing Data Numerically

Idea

It is often possible to summarize essential information about a sample numerically.

e.g.:

- How large? Location Parameters
- How variable? Dispersion Parameters

Already known from Boxplots

Location (How large?)

Median

Dispersion (How variable?)

Inter quartile range $(Q_3 - Q_1)$

3.1 Median and other Quartiles

The median is the 50% quantile of the data.

i.e.: half of the data are smaller or equal to the median, the other half are larger or equal.

The Quartiles

The first Quartile, Q_1: A quarter of the observations are smaller than or equal to Q_1 Three quarters are larger or equal.
i.e. Q_1 is the 25%-*Quantile*

The third Quartile, Q_3: Tree quarters of the observations are smaller than or equal to Q_3 One quarter are larger or equal.
i.e. Q_3 is the 75%-*Quantile*

3.2 Mean, Standard Deviation and Variance

Most frequently used

Location Parameter
The Mean $\overline{x}$

Dispersion Parameter
The Standard Deviation s

NOTATION:

Given data named $x_1, x_2, x_3, \ldots, x_n$
it is common to write $\overline{x}$ for the mean.

DEFINITION:

The mean of $x_1, x_2, \ldots, x_n$:

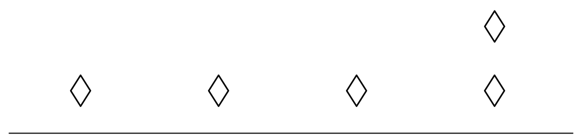
$$\begin{aligned} \overline{x} &= (x_1 + x_2 + \cdots + x_n)/n \\ &= \frac{1}{n}\sum_{i=1}^{n} x_i \end{aligned}$$

Geometric Interpretation of the Mean

Center of Gravity

Mean = Center of Gravity

Where is the center of gravity?



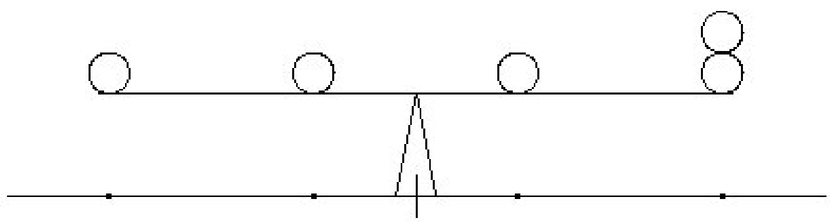
0 1 2 3

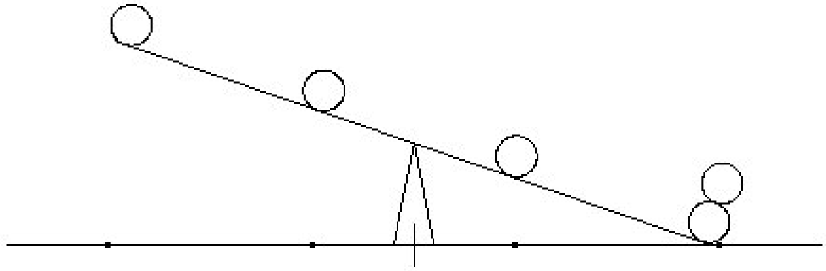
x

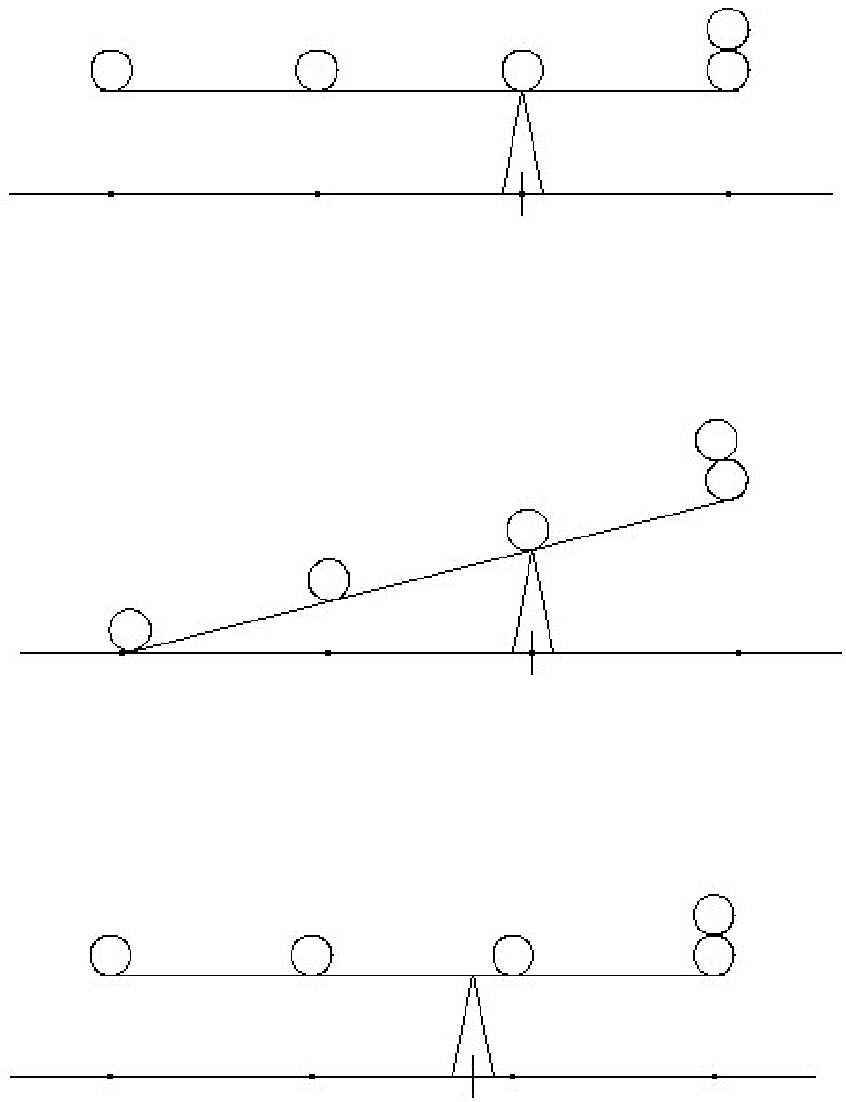
$m = 1.5$?

$m = 2$?

$m = 1.8$?







too small

too large

correct!

The Standard Deviation

How far do typical observations deviate from the mean?

The *Standard Deviation* σ ("sigma") is a slightly weired weighted mean of the deviations:

$$\sigma = \sqrt{\text{Sum(Deviations}^2)/n}$$

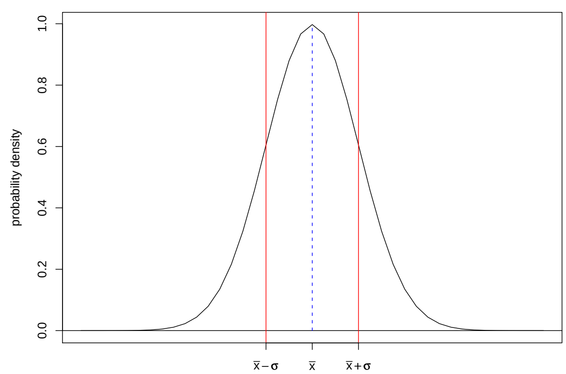
The formula for the *Standard Deviation* of $x_1, x_2, \ldots, x_n$:

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

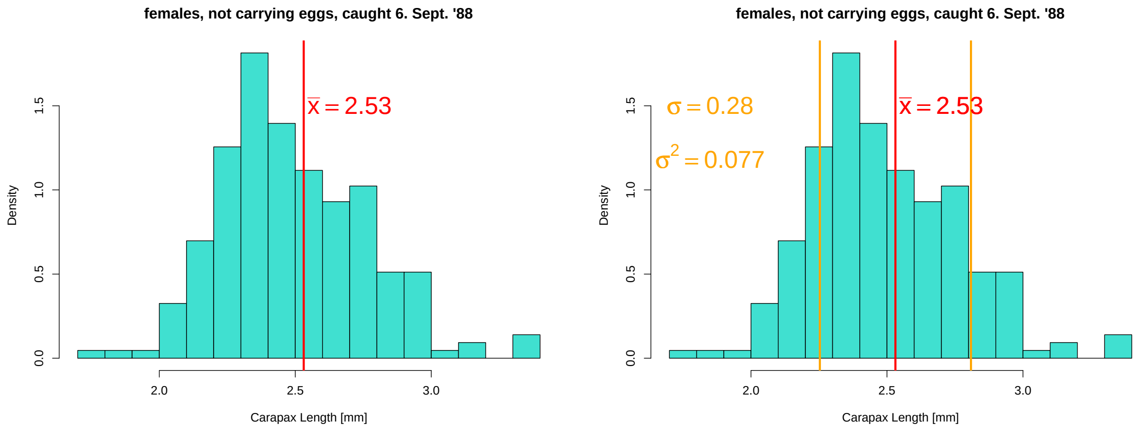
$\sigma^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^2$ is the *Variance.*

Rule of Thumb for the Standard Deviation

In more or less bell-shaped (i.e. single peak, symmetic) distributions: ca. 2/3 are located between $\overline{x} - \sigma$ und $\overline{x} + \sigma$.



Standard Deviation of Carapace lengths from 6.9.88



In this case 72% are between $\overline{x} - \sigma$ and $\overline{x} + \sigma$

Variance of Carapace lengths from 6.9.88

All Carace Lengths in North Sea: $\mathcal{X} = (X_1, X_2, \ldots, X_N)$.Carapace Length in our Sample: $\mathcal{S} = (S_1, S_2, \ldots, S_{n=215})$Sample Variance:

$$\sigma_{\mathcal{S}}^2 = \frac{1}{n} \sum_{i=1}^{215} (S_i - \overline{S})^2 \approx 0.0768$$

Can we use 0.0768 as estimation for $\sigma_{\mathcal{X}}^2$, the variance in the whole population?Yes, we can! However, $\sigma_{\mathcal{S}}^2$ is on average by a factor of $\frac{n-1}{n}$ $(= 214/215 \approx 0.995)$ smaller than $\sigma_{\mathcal{X}}^2$.

Variances

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

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

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

Usually, "Standard Deviation (SD) of $\mathcal{S}$" refers to the corrected s.

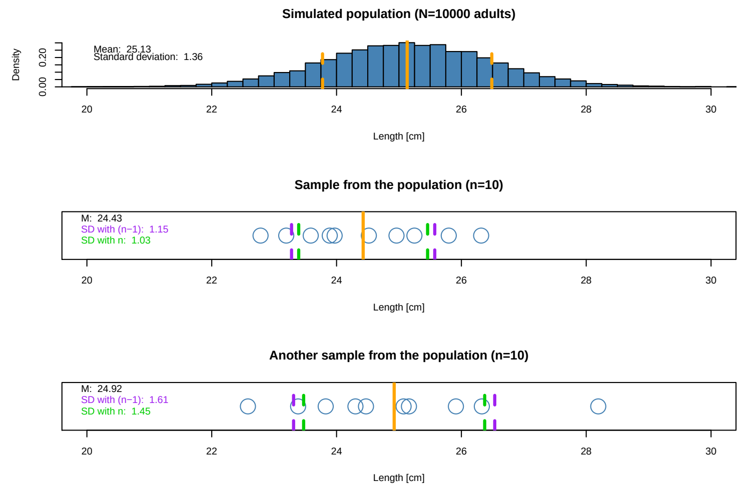
Example: Computing SD

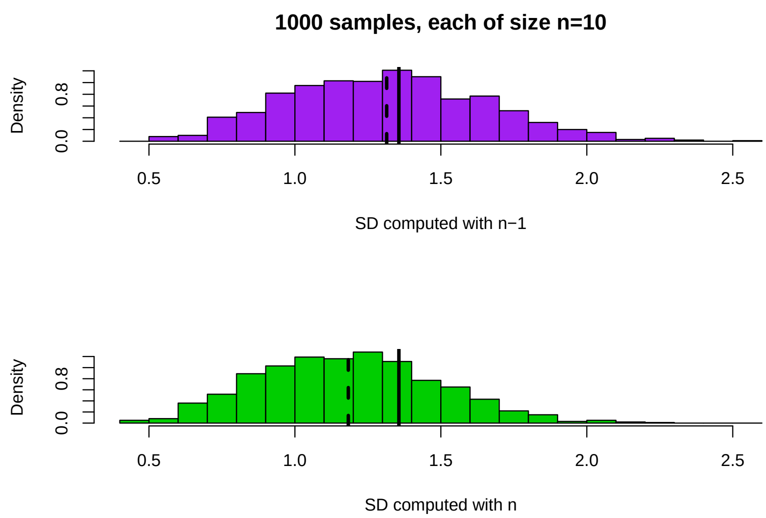
Given Data $\overline{x} = ?$ $\overline{x} = 10/5 = 2$

						$\sum$
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

$$
\begin{aligned}
s^2 &= \left(\sum_x (x-\overline{x})^2\right)/(n-1) \\
&= 16/(5-1) = 4 \\
s &= 2
\end{aligned}
$$

3.2.1 Computing σ with n or $n-1$?





Computing σ with n or $n-1$?

The standard deviation σ of a random variable with n equally probable outcomes $x_1, \ldots, x_n$ (z.B. rolling a dice) is clearly defined by

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

If $x_1, \ldots, x_n$ is a sample (the usual case in statistics) you should rather use the formula

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

4 When may mean values and standard deviation be misleading?

Mean and SD...

- characterize data well if the distribution is bell-shaped
- and must be interpreted with caution in other cases

We will exemplify this with textbook examples from ecology, see e.g.

References

[BTH08] M. Begon, C. R. Townsend, and J. L. Harper. *Ecology: From Individuals to Ecosystems.* Blackell Publishing, 4 edition, 2008.

When original data were not available, we generated similar data sets by computer simulation. So do not believe all data points.

4.0.1 example: picky wagtails

Wagtails eat dung flies

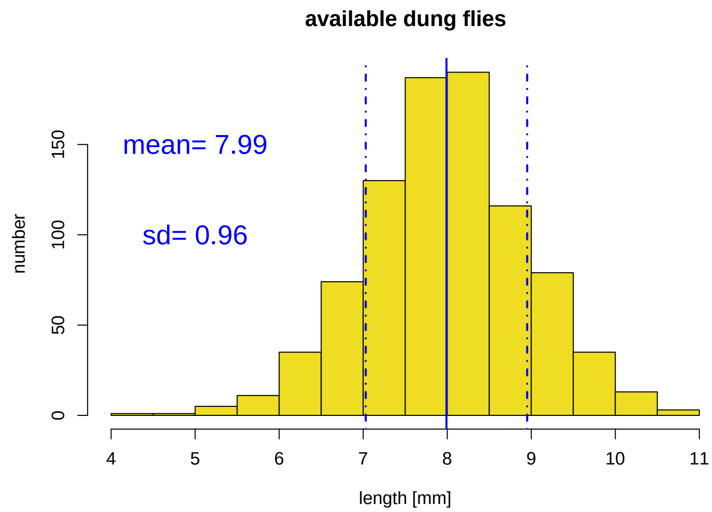
Predator	Prey
White Wagtail *Motacilla alba alba*	Dung Fly *Scatophaga stercoraria*

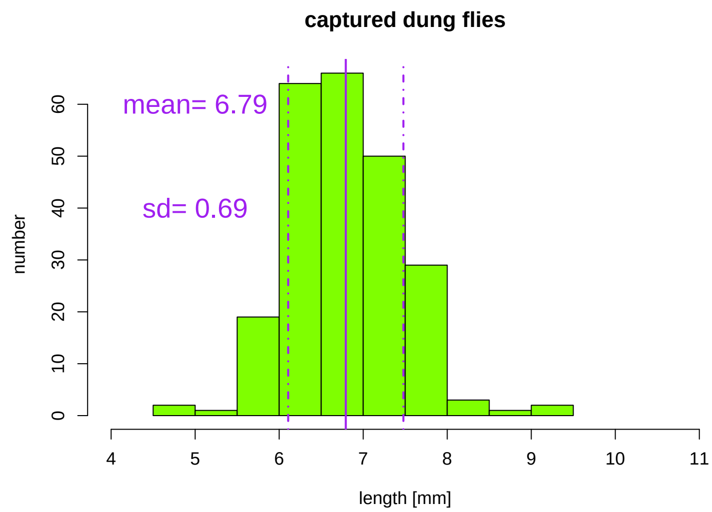
Conjecture

- Size of flies varies.
- efficiency for wagtail = energy gain / time to capture and eat
- lab experiments show that efficiency is maximal when flies have size 7mm

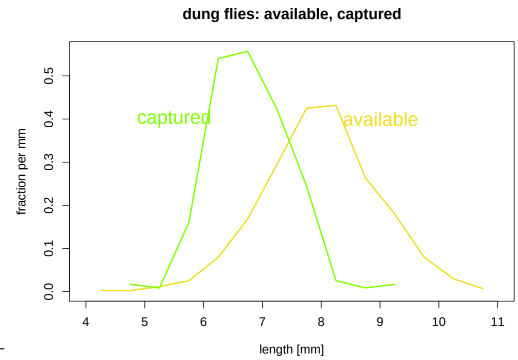
References

[Dav77] N.B. Davies. Prey selection and social behaviour in wagtails (Aves: Motacillidae). *J. Anim. Ecol.*, 46:37–57, 1977.





numerical comparison of size distributions



	captured		available
mean	6.29	<	7.99
sd	0.69	<	0.96

Interpretation

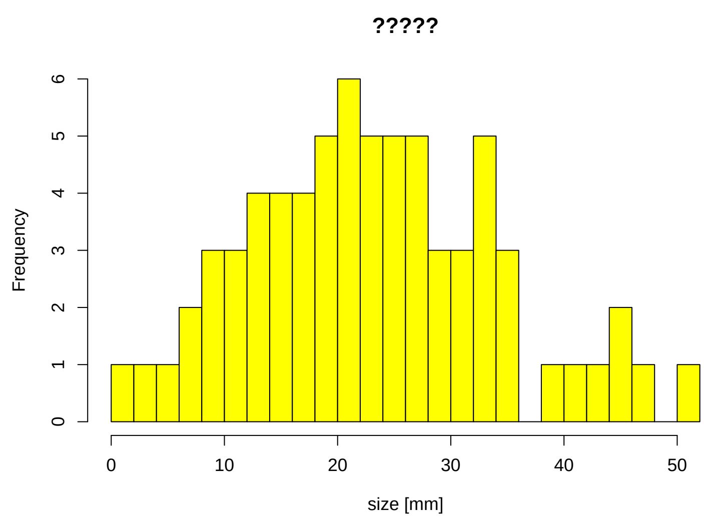
The birds prefer dung-flies from a relatively narrow range around the predicted optimum of 7mm.

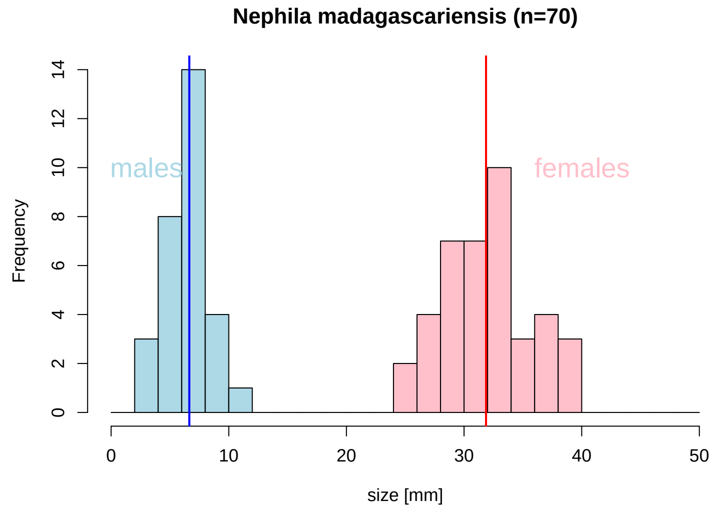
The distributions in this example were bell-shaped, and the 4 numbers (means and standard deviations) were appropriate to summarize the data.

4.0.2 example: spider men & spider women

Nephila madagascariensis
image (c) by Bernard Gagnon

Simulated Data:
70 sampled spiders
mean size: 21.05 mm
sd of size :12.94 mm





Conclusion from spider example

If data comes from different groups, it may be reasonable to compute mean an sd separately for each group.

4.0.3 example: copper-tolerant browntop bent

Copper Tolerance in Browntop Bent

Browntop Bent	Copper
Agrostis tenuis	*Cuprum*
image (c) Kristian Peters	Hendrick met de Bles

References

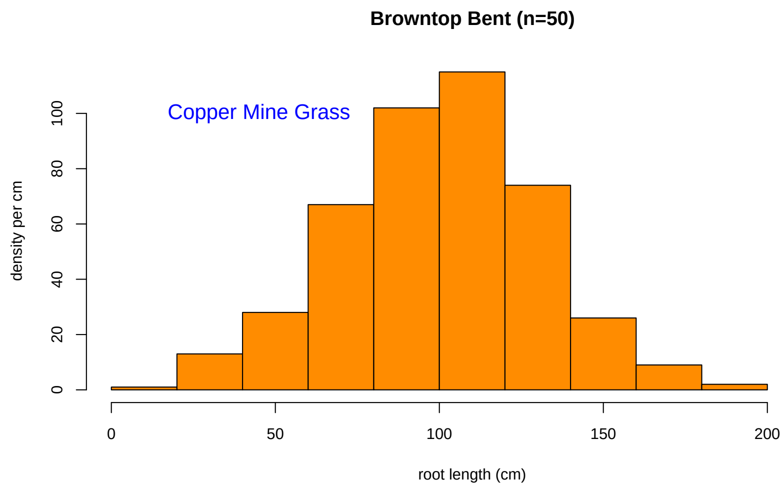
[Bra60] A.D. Bradshaw. Population Differentiation in *agrostis tenius Sibth.* III. populations in varied environments. *New Phytologist*, 59(1):92 – 103, 1960.

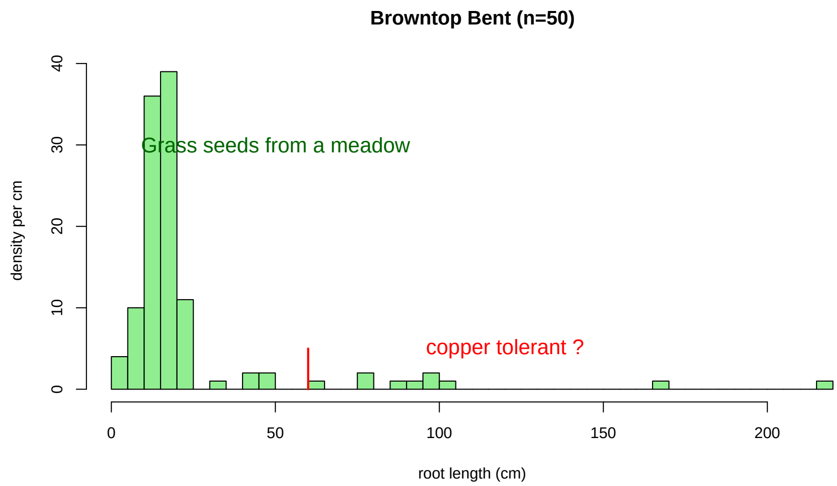
[MB68] T. McNeilly and A.D Bradshaw. Evolutionary Processes in Populations of Copper Tolerant Agrostis tenuis Sibth. *Evolution*, 22:108–118, 1968.

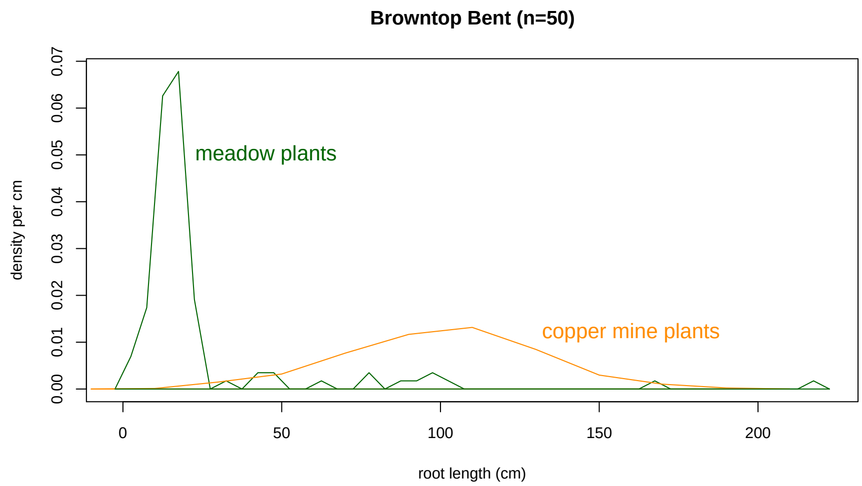
Again, we have no access to original data and use simulated data.

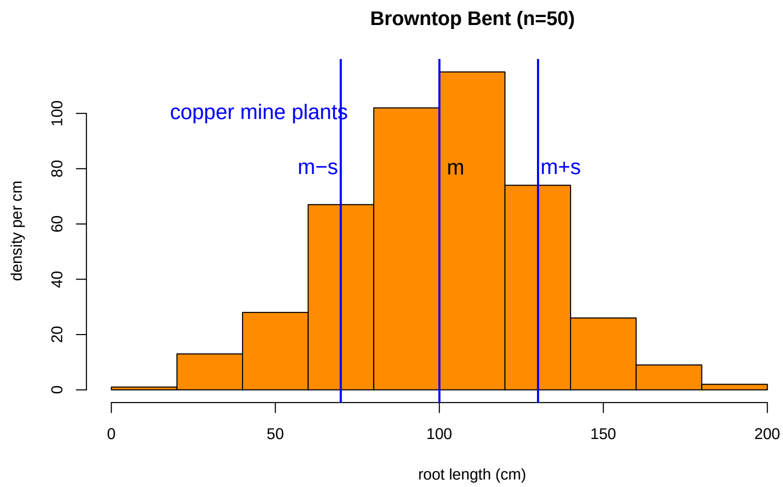
Adaptation to copper?

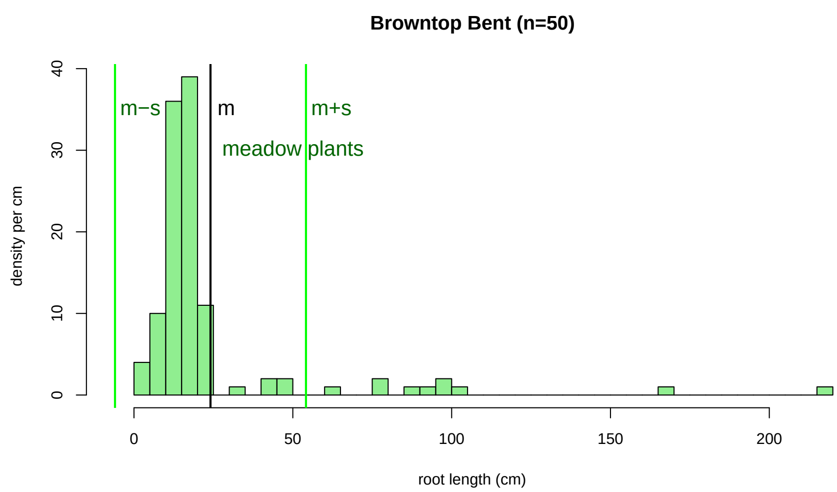
- root length indicates copper tolerance
- measure root lengths of plants near copper mine
- take seeds from clean meadow and sow near copper mine
- measure root length of these “meadow plants” in copper environment

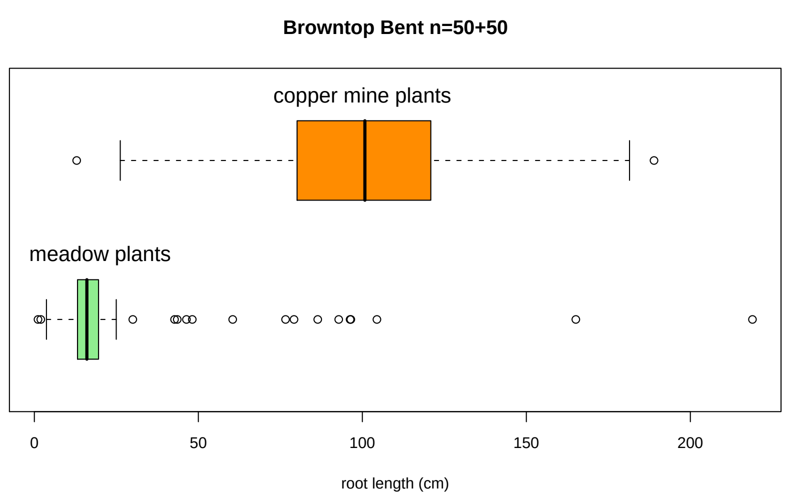

Browntop Bent (n=50)
Copper Mine Grass
density per cm
0
20
40
60
80
100
0
50
100
150
200
root length (cm)


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)







2/3 of the data within [m-sd,m+sd]???? No!

quartiles of root length [cm]

	min	Q_1	median	Q_3	max
copper adapted	12.9	80.1	100.8	120.9	188.9
from meadow	1.1	13.2	16.0	19.6	218.9

Conclusion from browntop bent example

Sometimes the two numbers
m and sd
give not enough information.

In this example the five quartiles
min, Q_1, median, Q_3, max
that are shown in the boxplot are more appropriate.

Conclusions from this section

Always visually inspect the data!

Never rely on summarising values alone!

Image copyright notes see

http://en.wikipedia.org/wiki/File:Motacilla_alba_alba.JPG [Artur Mikołajewski] http://commons.wikimedia.org/wiki/File:Scatophaga_stercoraria_1_Luc_Viatour.jpg [Viatour Luc] http://commons.wikimedia.org/wiki/File:N\unhbox\voidb@x\bgroup\let\unhbox\voidb@x\setbox\@tempboxa\hbox{e\global\mathchardef\accent@spacefactor\spacefactor}\accent19e\egroup\spacefactor\accent@spacefactorphila_inaurata_Madagascar_02.jpg [Bernard Gagnon] http://de.wikipedia.org/w/index.php?title=Datei:Agrostis_capillaris.jpeg [Kristian Peters] http://de.wikipedia.org/w/index.php?title=Datei:Hendrick_met_de_Bles_001.jpg [Hendrick met de Bles]

Some of the things you should be able to explain

- How to study for this course
- what is a density
- how to interpret histograms and density plots
- boxplots and stripcharts and when to use them
- quartiles and median
- mean and sd and how to guess them from histograms, density plots, stripcharts or scatterplots
- var and sd: when to divide by $n - 1$ and why
- why visualizing data and when means etc. can be misleading