

Statistics for EES
General Introduction and Descriptive Statistics

Dirk Metzler

April 7, 2026

Contents

Contents

1	Intro: What is Statistics?	1
2	Data Visualization	3
2.1	Histograms und Density Polygons	5
2.1.1	Histograms: Densities or Numbers?	10
2.2	Stripcharts and Boxplots	10
2.3	Conclusions	17
2.4	Pie charts or bar charts? An experiment	17
2.5	Example: blue area challenge results from 2018	18
3	Summarizing Data Numerically	20
3.1	Median and other Quartiles	20
3.2	Mean, Standard Deviation and Variance	21
3.2.1	Computing σ with n or $n - 1$?	26
4	When may mean values and standard deviation be misleading?	27
4.0.1	example: spider sizes	27

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

There is sampling variation

There is variation in measurements

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 principles 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 and 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

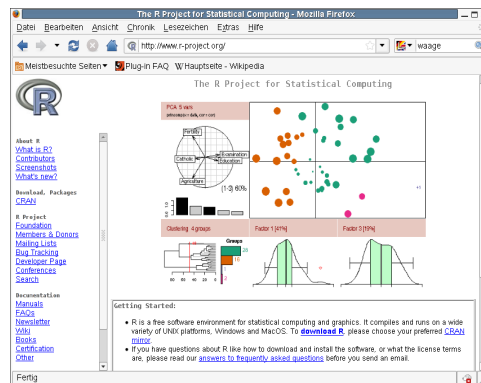
What you need to answer the questions:

- 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

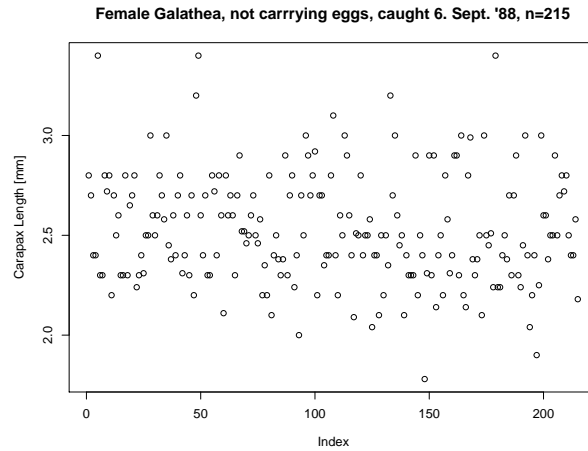
Galathea intermedia

Squat Lobsters, caught 6. Sept 1988

Helgoländer Tiefe Rinne, North Sea

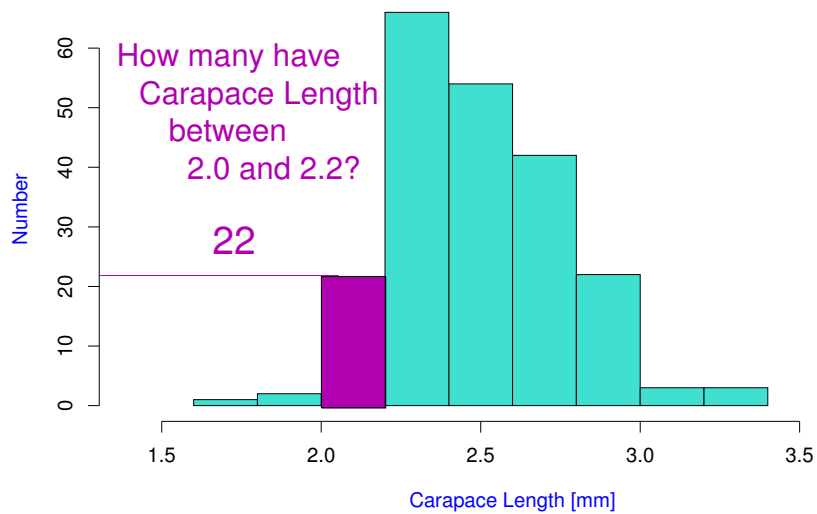
Carpace Lengths (mm): Females, not egg-carrying ($n = 215$)

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

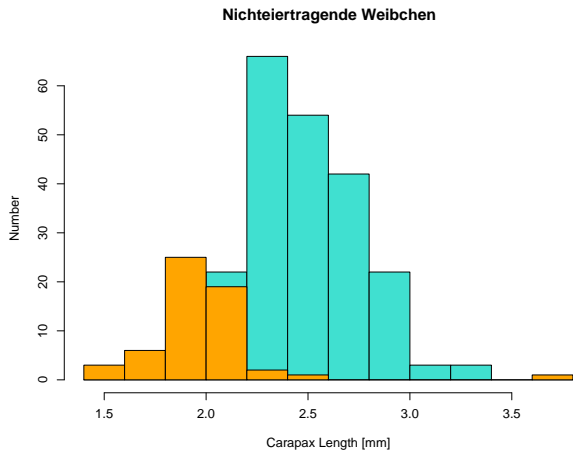


2.1 Histograms und Density Polygons

Female Galathea, not egg-carrying, caught 6. Sept. '88, n=215



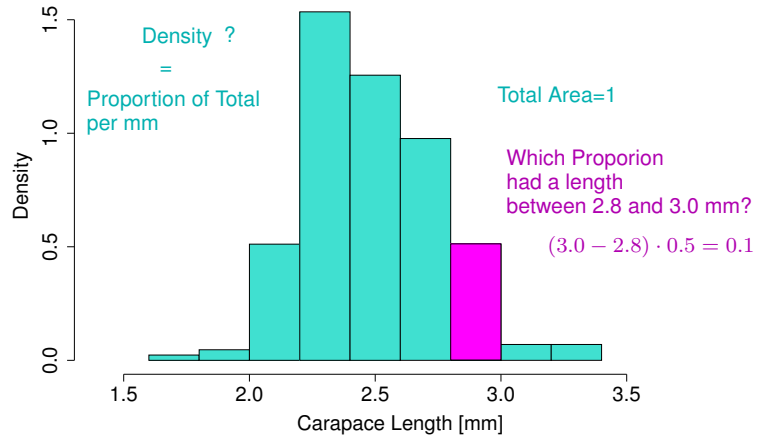
Comparing the two Distributions



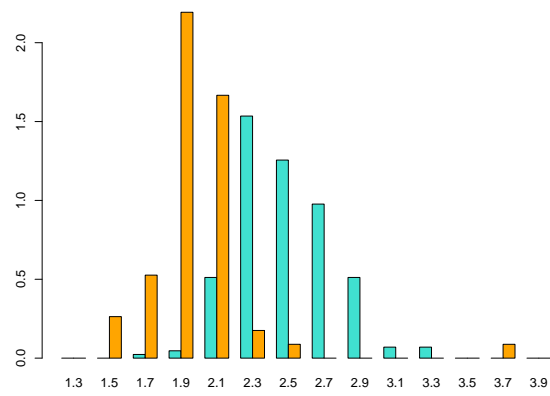
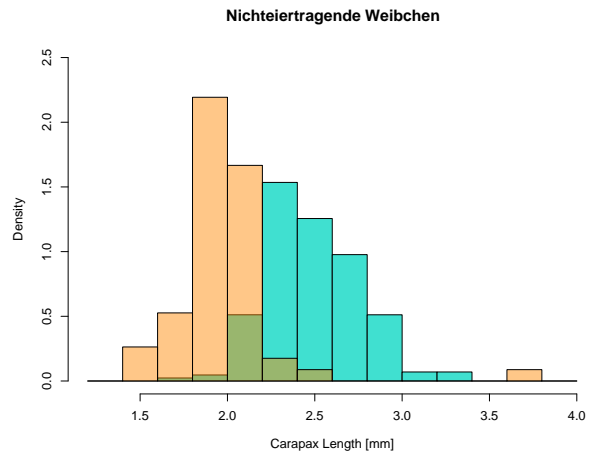
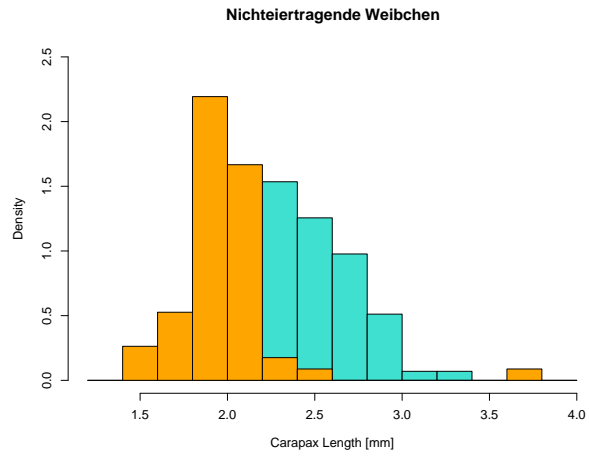
Problem: different sample sizes
 6.9.1988 : $n = 215$
 3.11.1988 : $n = 57$

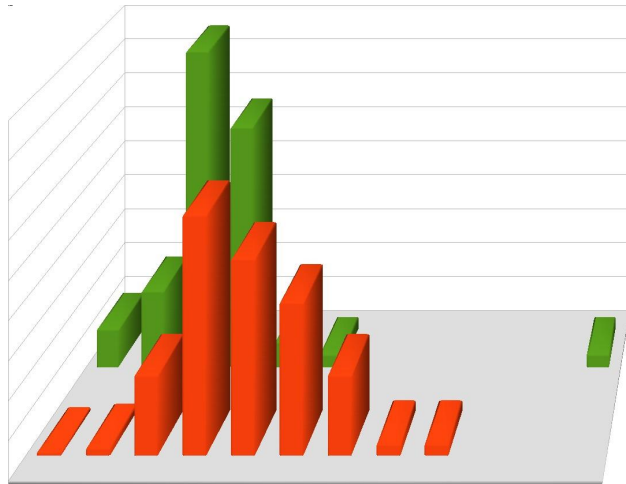
Idea: scale y-axis such that each distribution has total area 1.

Female Crabs, not egg-carrying, caught 6. Sept. '88, n=215



How to compare the two distributions?





My Advice

If you are a commercial artist:

Impress everybody with cool 3D graphics!

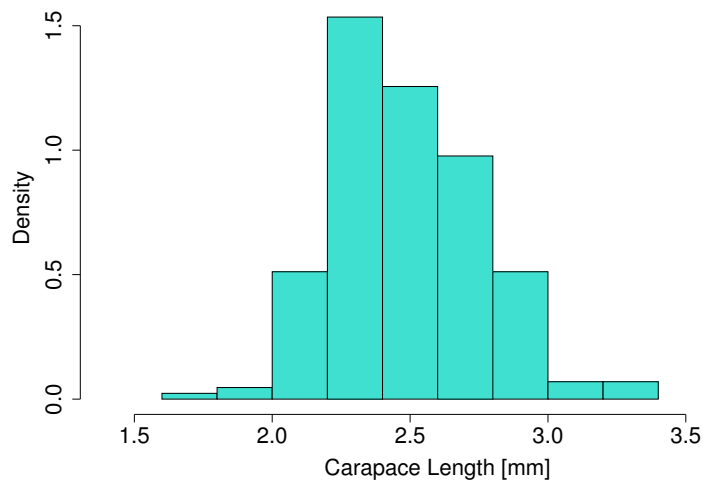
If you are a scientist:

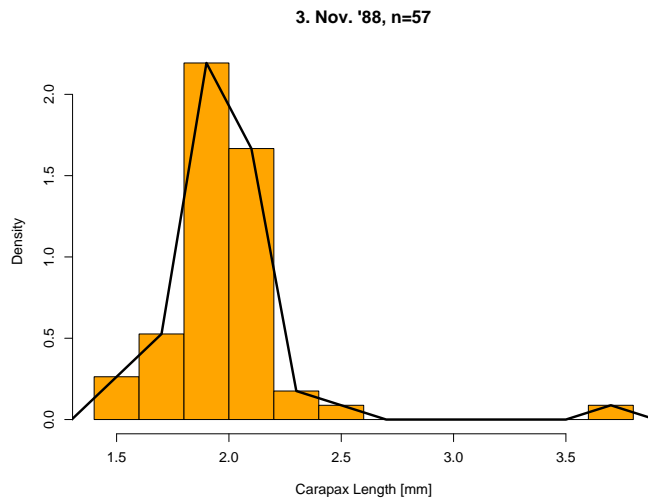
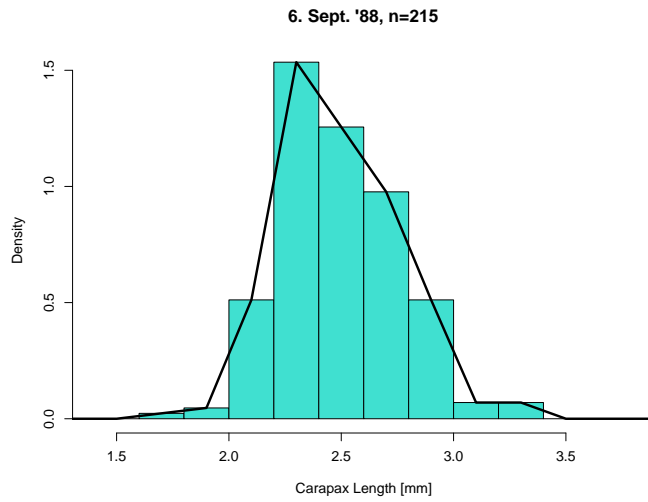
Visualize your data in clear and simple 2D plots.

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

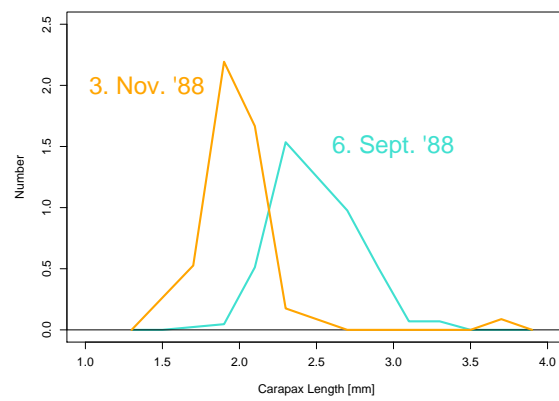
Simple and Clear: Density Polygons

Female Crabs, not egg-carrying, caught 6. Sept. '88, n=215





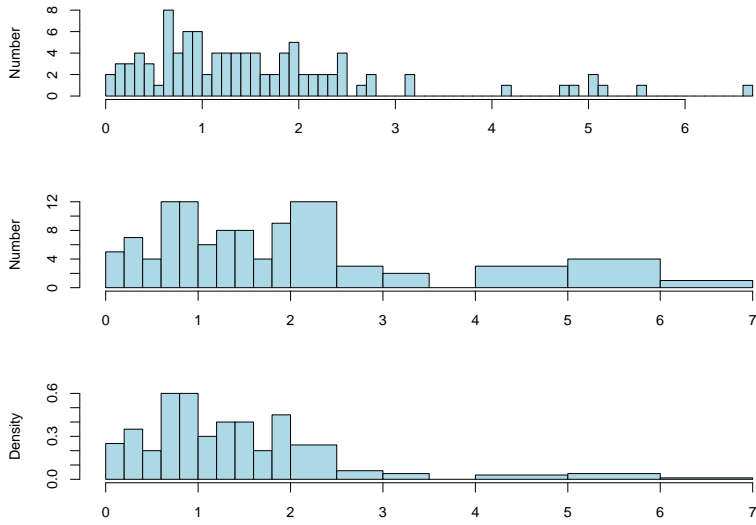
Convenient to show two or more Density Polygons in one plot



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

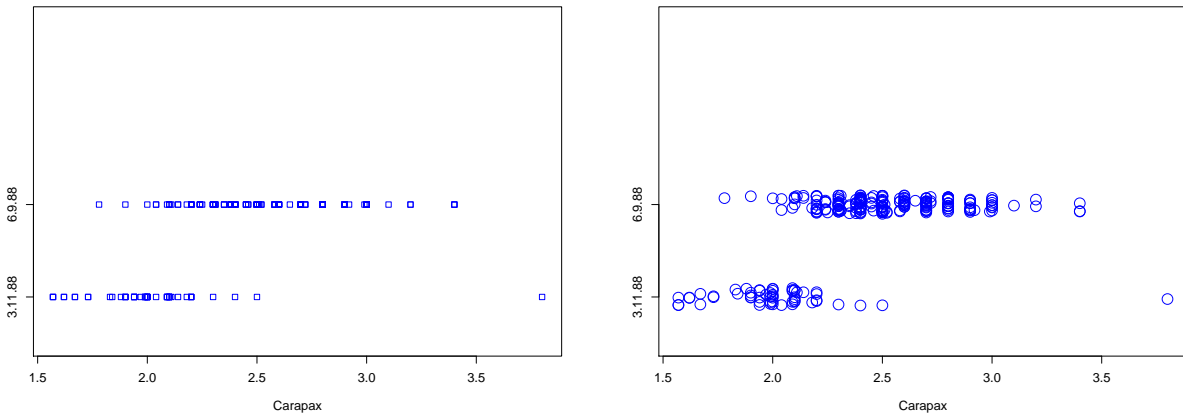
2.1.1 Histograms: Densities or Numbers?

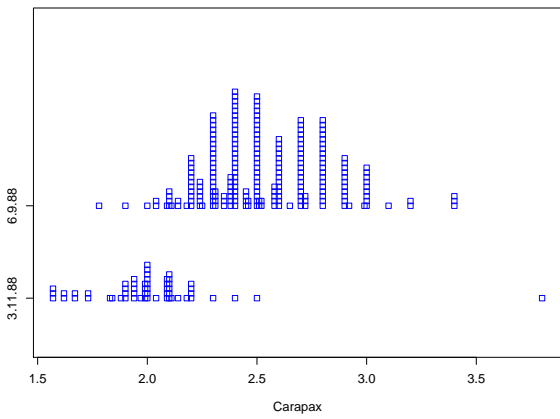
Number vs. Density



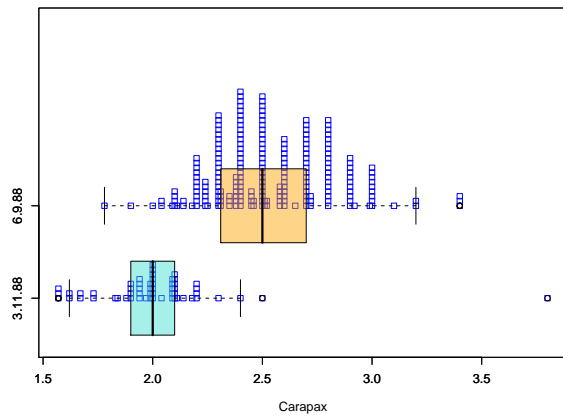
Histograms with unequal intervals should show densities, not numbers!

2.2 Stripcharts and Boxplots

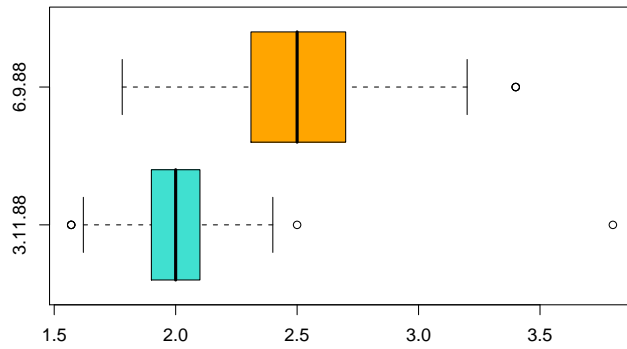




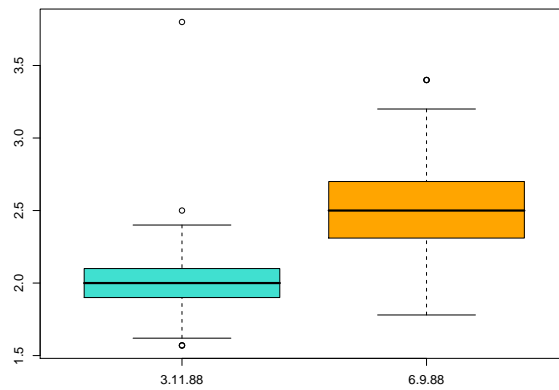
Stripchart + Boxplots, horizontal



Boxplots, horizontal



Boxplots, vertikal

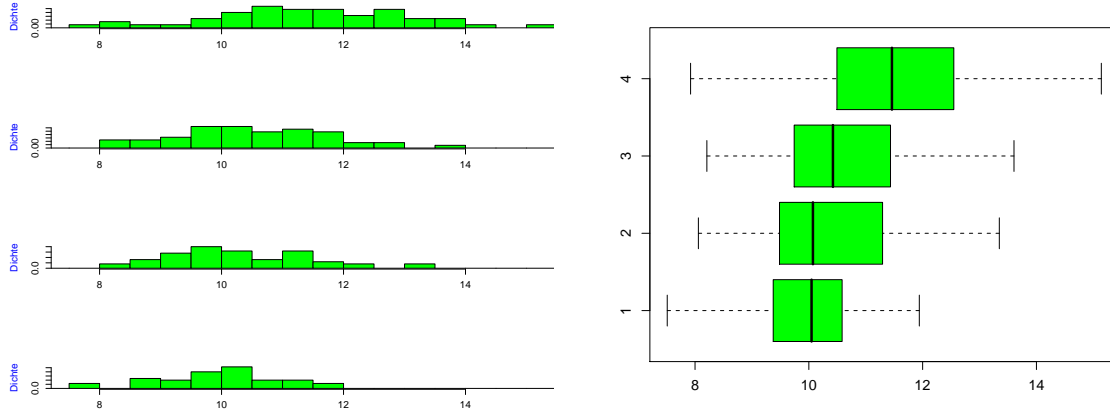


Simplify to understand

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

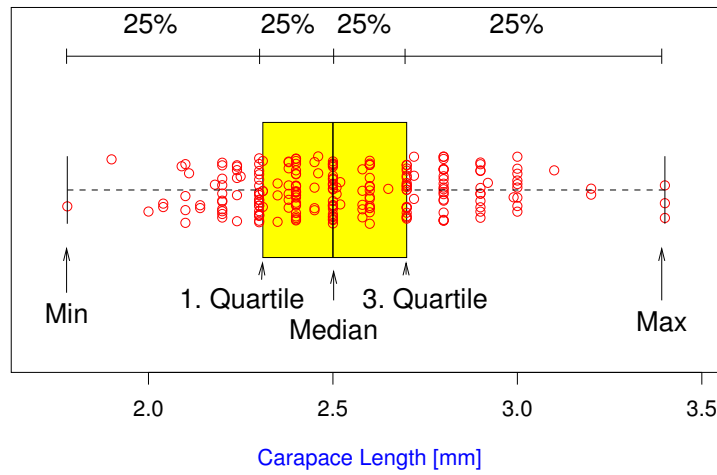
Sometimes too comprehensive.

Comparison of four groups

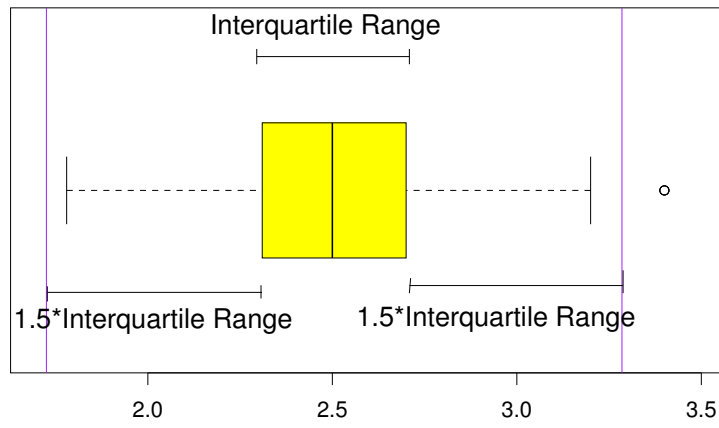


The Boxplot

Boxplot, simple type

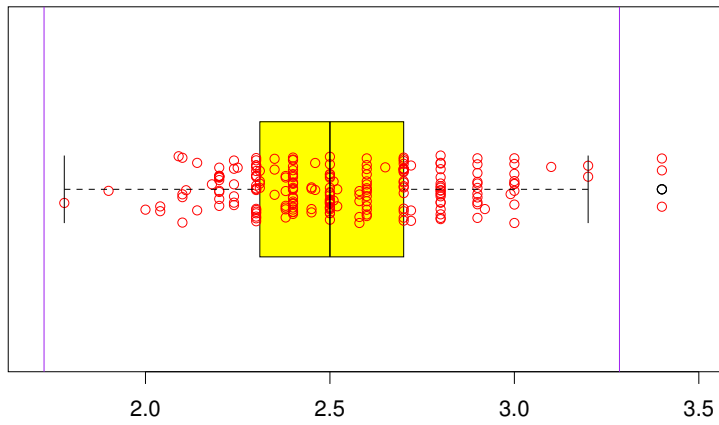


Boxplot, Standard Type



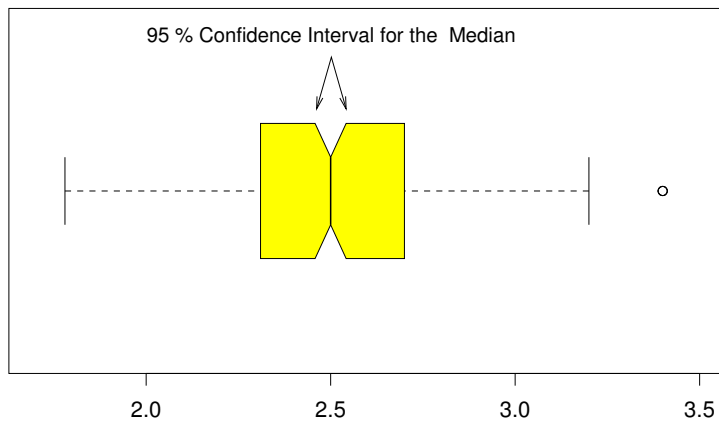
Carapace Length [mm]

Boxplot, Standard Type



Carapace Length [mm]

Boxplot Professional



Carapace Length [mm]

Example: Darwin Finches

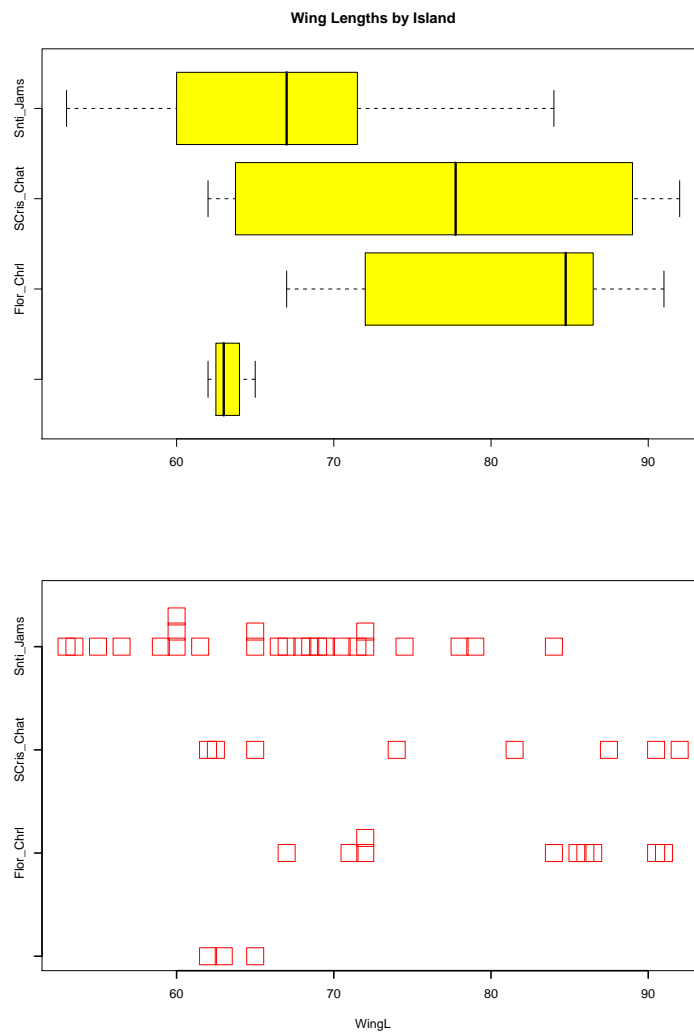
Darwin's collection of Finches

References

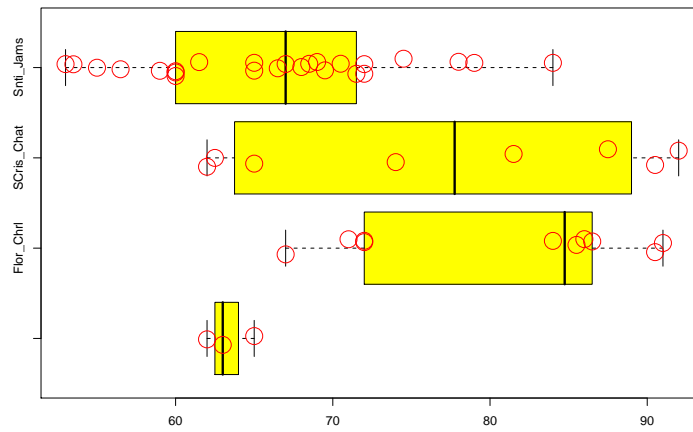
[1] Sulloway, F.J. (1982) The Beagle collections of Darwin's Finches (Geospizinae). *Bulletin of the British Museum (Natural History), Zoology series* **43**: 49-94.

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

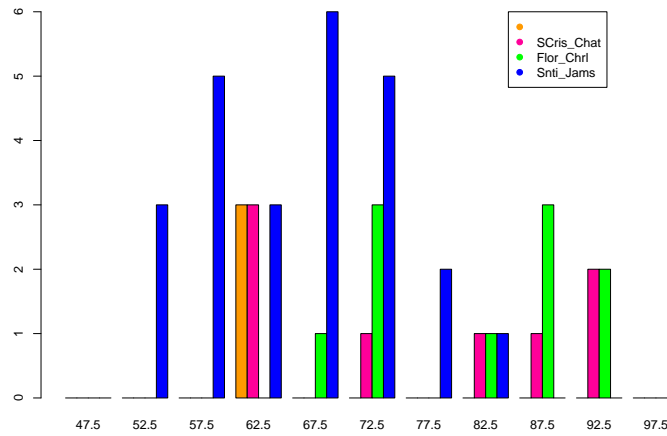
Wing Sizes of Darwin's Finches



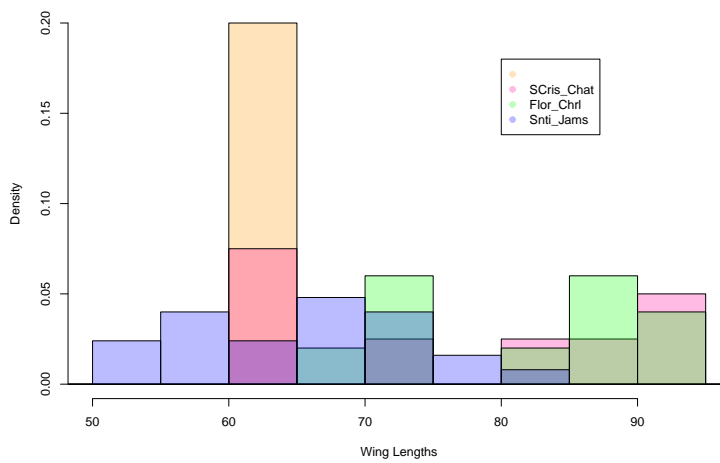
Wing Lengths by Island

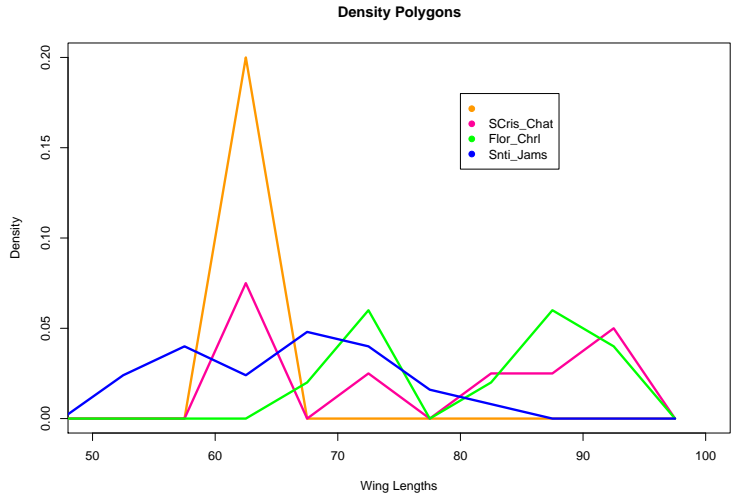


Barplot of Wing Lengths (Numbers)

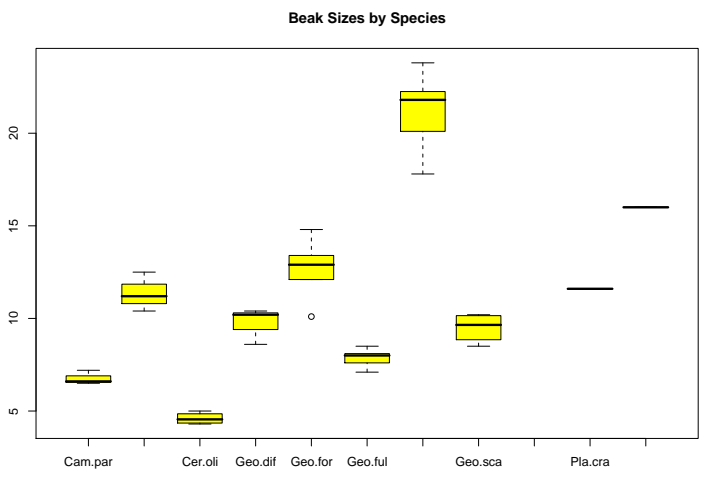


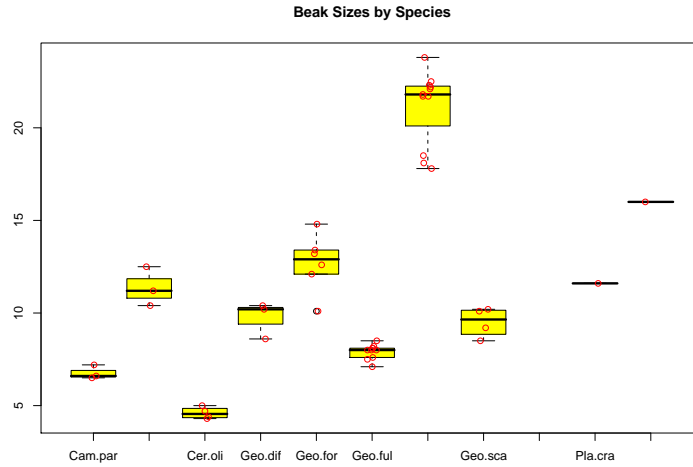
Histogramm (Densities!) with transparen colors





Beak Sizes of Darwin's Finches



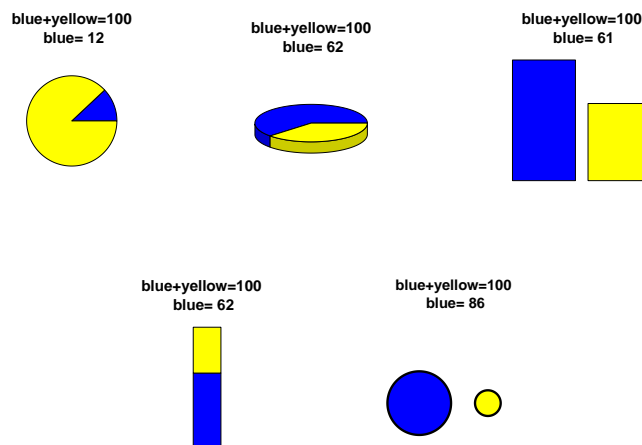


2.3 Conclusions

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-transparent 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

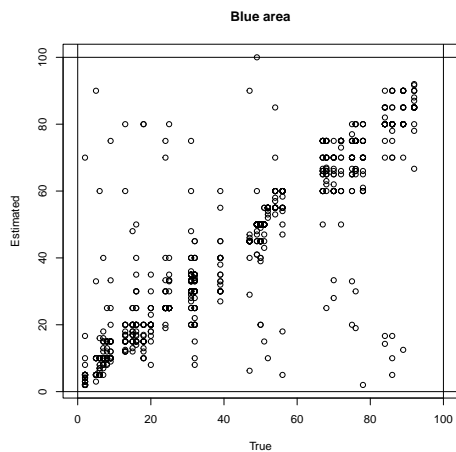
```
est <- read.csv("DataAndR/bluearea_estimates_2018.csv")
str(est)

## '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
```

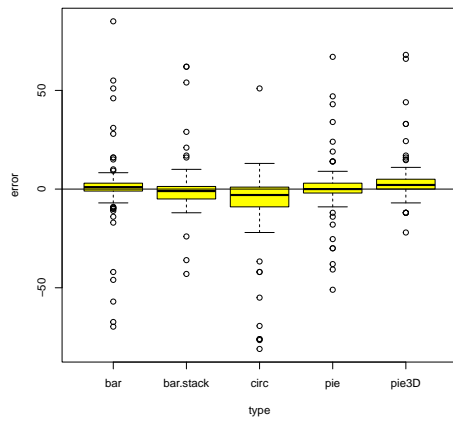
```
plot(est$true,est$estimate,xlab="True", ylab="Estimated",
      main="Blue area",xlim=c(0,100))
abline(h=c(0,100))
abline(v=c(0,100))
```



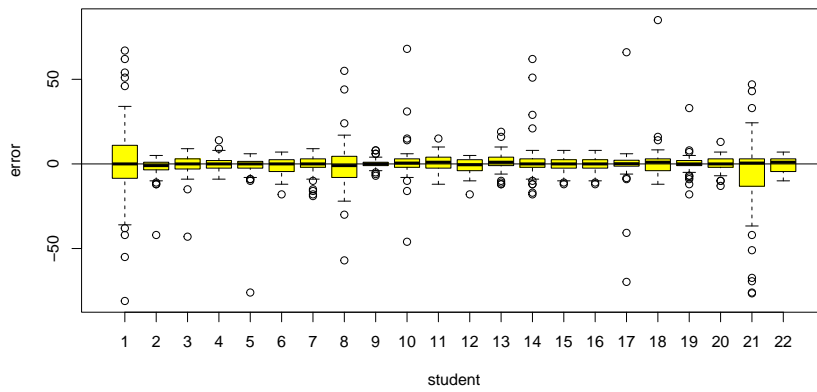
```
est$error <- est$estimate-est$true
str(est)

## '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 ...
```

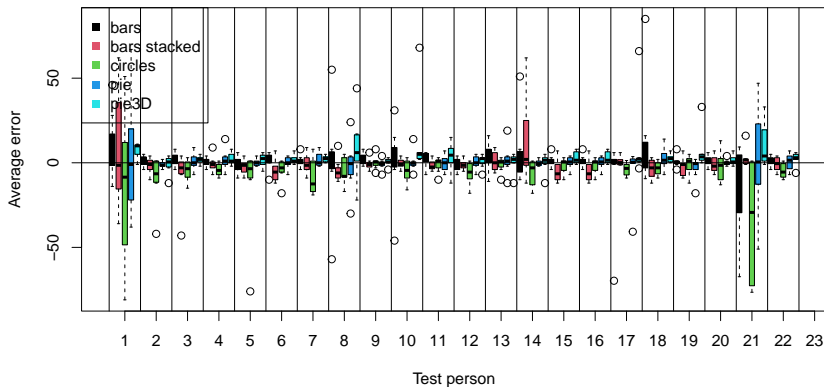
```
boxplot(error~type,est,col="yellow")
abline(h=0)
```



```
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 \bar{x}

Dispersion Parameter

The Standard Deviation s

NOTATION:

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

DEFINITION:

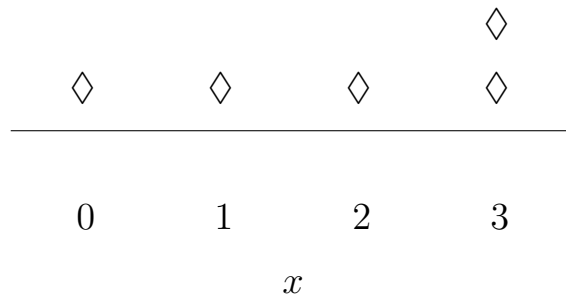
The mean of x_1, x_2, \dots, x_n :

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

Geometric Interpretation of the Mean
Center of Gravity

Mean = Center of Gravity

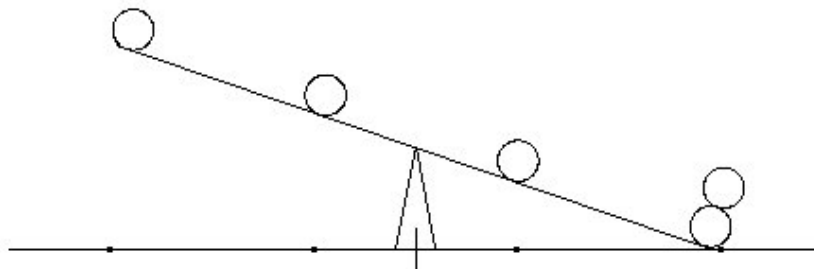
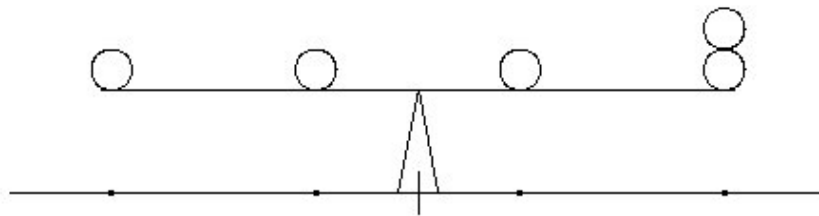
Where is the center of gravity?

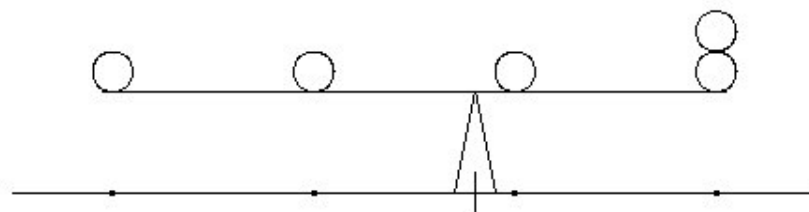
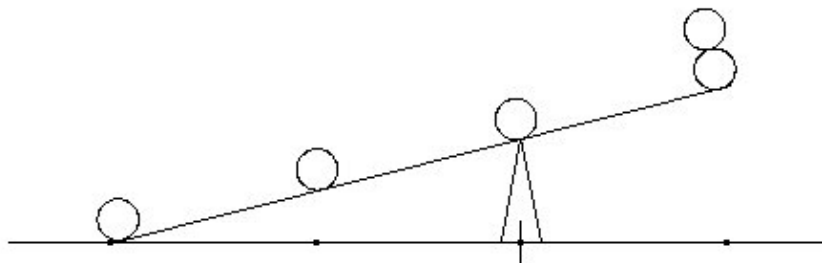
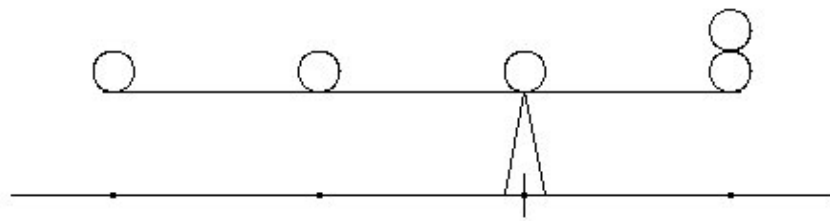


$$m = 1.5 ?$$

$$m = 2 ?$$

$$m = 1.8 ?$$





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 weird weighted mean of the deviations:

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

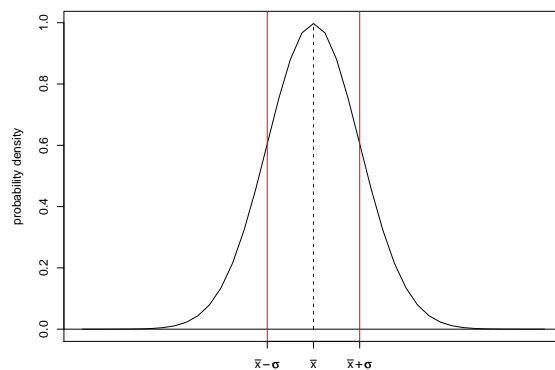
The formula for the *Standard Deviation* of x_1, x_2, \dots, x_n :

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

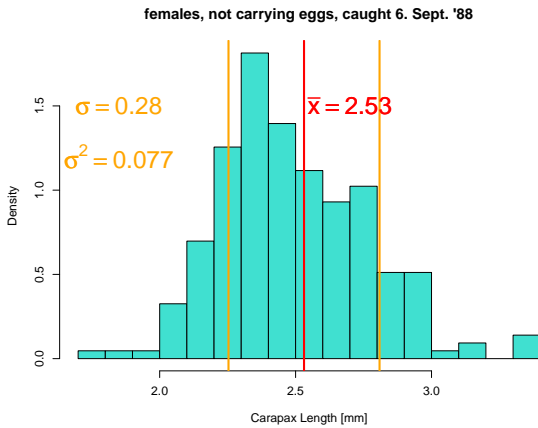
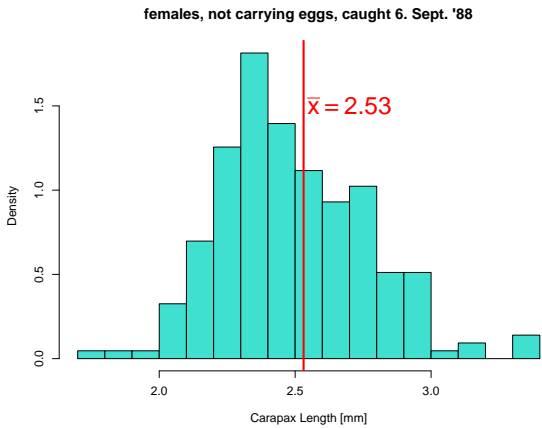
$\sigma^2 = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2$ is the *Variance*.

Rule of Thumb for the Standard Deviation

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



Standard Deviation of Carapace lengths from 6.9.88



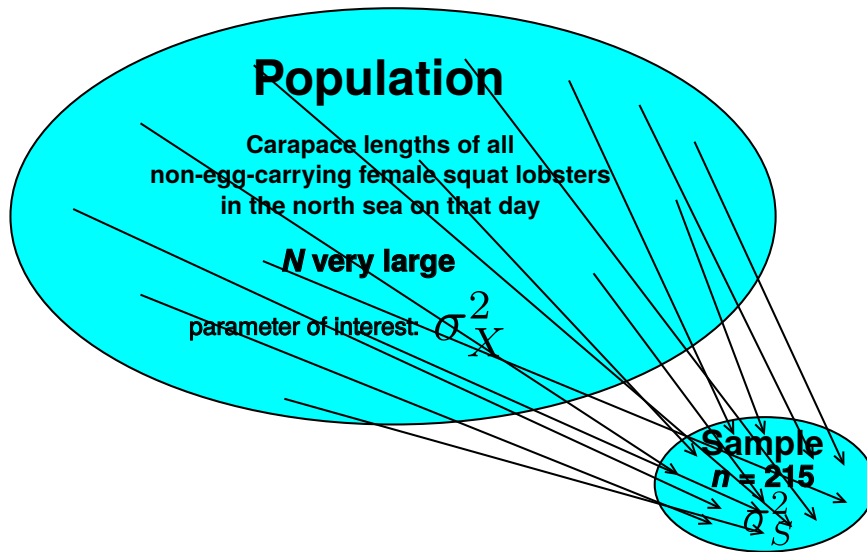
In this case 72% are between $\bar{x} - \sigma$ and $\bar{x} + \sigma$

Variance of Carapace lengths from 6.9.88

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

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

Can we use 0.0768 as estimation for $\sigma_{\mathcal{X}}^2$, the variance in the whole population?



Variance of Carapace lengths from 6.9.88

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

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

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

Variations

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

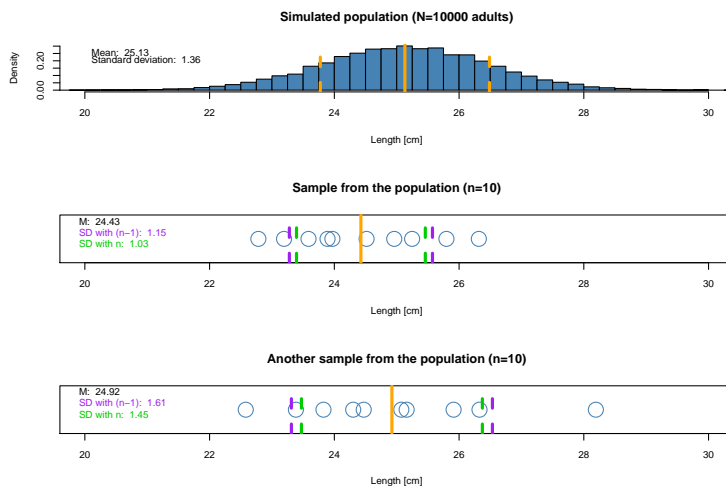
Sample Variance: $\sigma_S^2 = \frac{1}{n} \sum_{i=1}^n (S_i - \bar{S})^2$

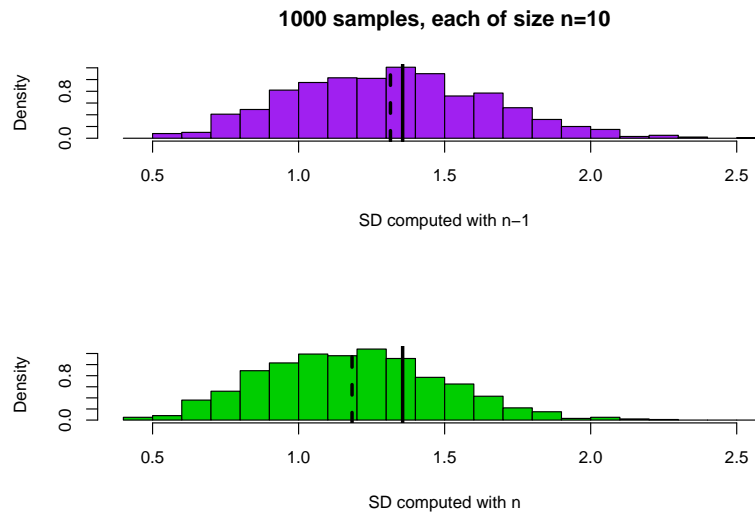
(Corrected) Sample Variance:

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

Usually, “Standard Deviation (SD) of S ” refers to the corrected s .

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, \dots, x_n (z.B. rolling a dice) is clearly defined by

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

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

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

4 When may mean values and standard deviation be misleading?

Mean and SD...

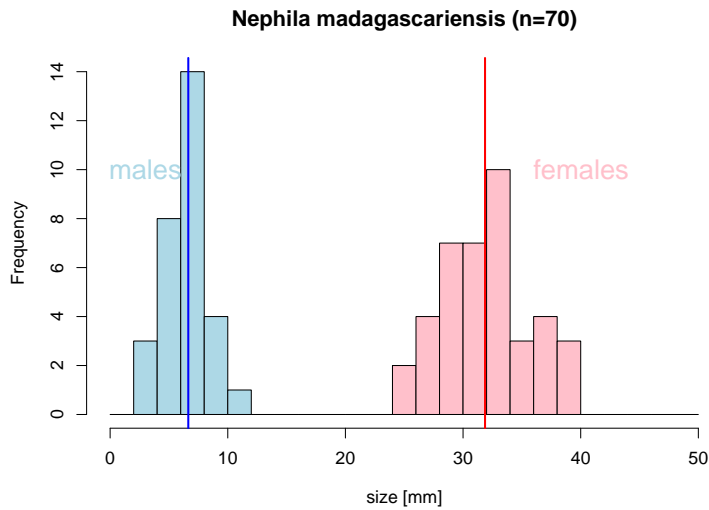
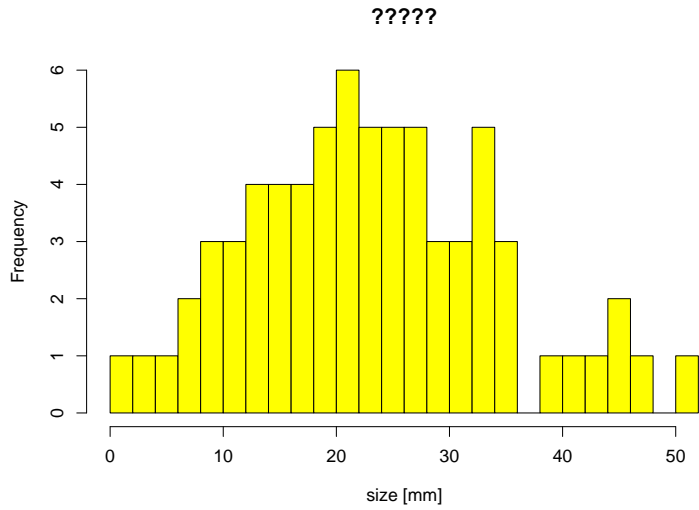
- characterize data well if the distribution is bell-shaped
- and must be interpreted with caution in other cases

4.0.1 example: spider sizes

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 and sd separately for each group.

Conclusions from this section

Always visually inspect the data!

Never rely on summarising values alone!

Image copyright notes see

http://commons.wikimedia.org/wiki/File:N:\unhbox\voidb@x\bgroup\let\unhbox\voidb@x\setbox\@tempboxa\hbox{e\global\mathchardef\accent@spacefactor\spacefactor}\let\begin\group\end\group\relax\let\ignorespaces\relax\accent19e\egroup\spacefactor\accent@spacefactor\phila_inaurata_Madagascar_02.jpg [Bernard Gagnon]

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