Statistics for EES
Introduction to R and Descriptive Statistics

Dirk Metzler

April 18, 2016

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?

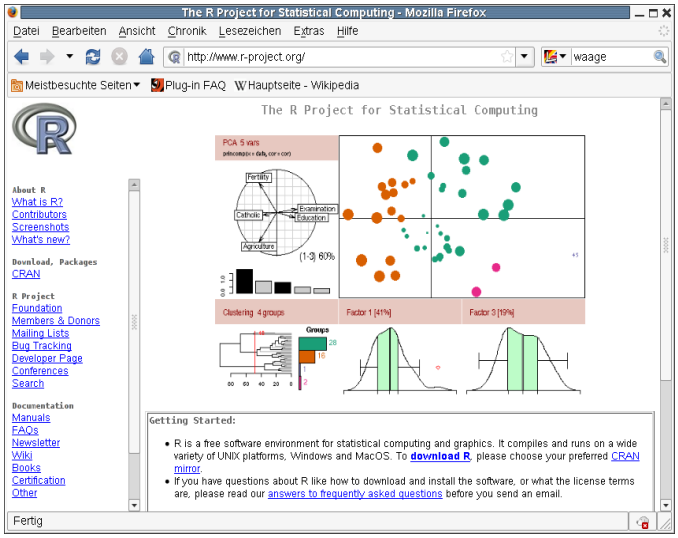
Nature is full of Variability
How to make sense of variable data?
Use mathematical theory of randomness:[0.5ex] *Probability*.

Statistics = *Data Analysis* based on *Probabilistic Models*

Descriptive Statistics

Descriptive Statistcs is the first look at the data.

Statistics Software R



http://www.r-project.org

2 Data Visialization

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

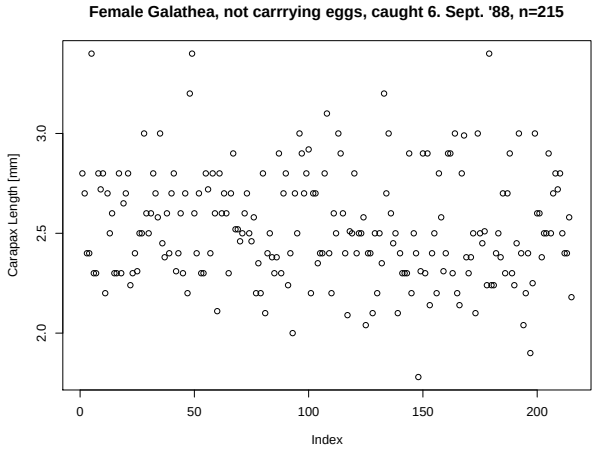
Galathea intermedia

Squat Lobsters, caught 6. Sept 1988

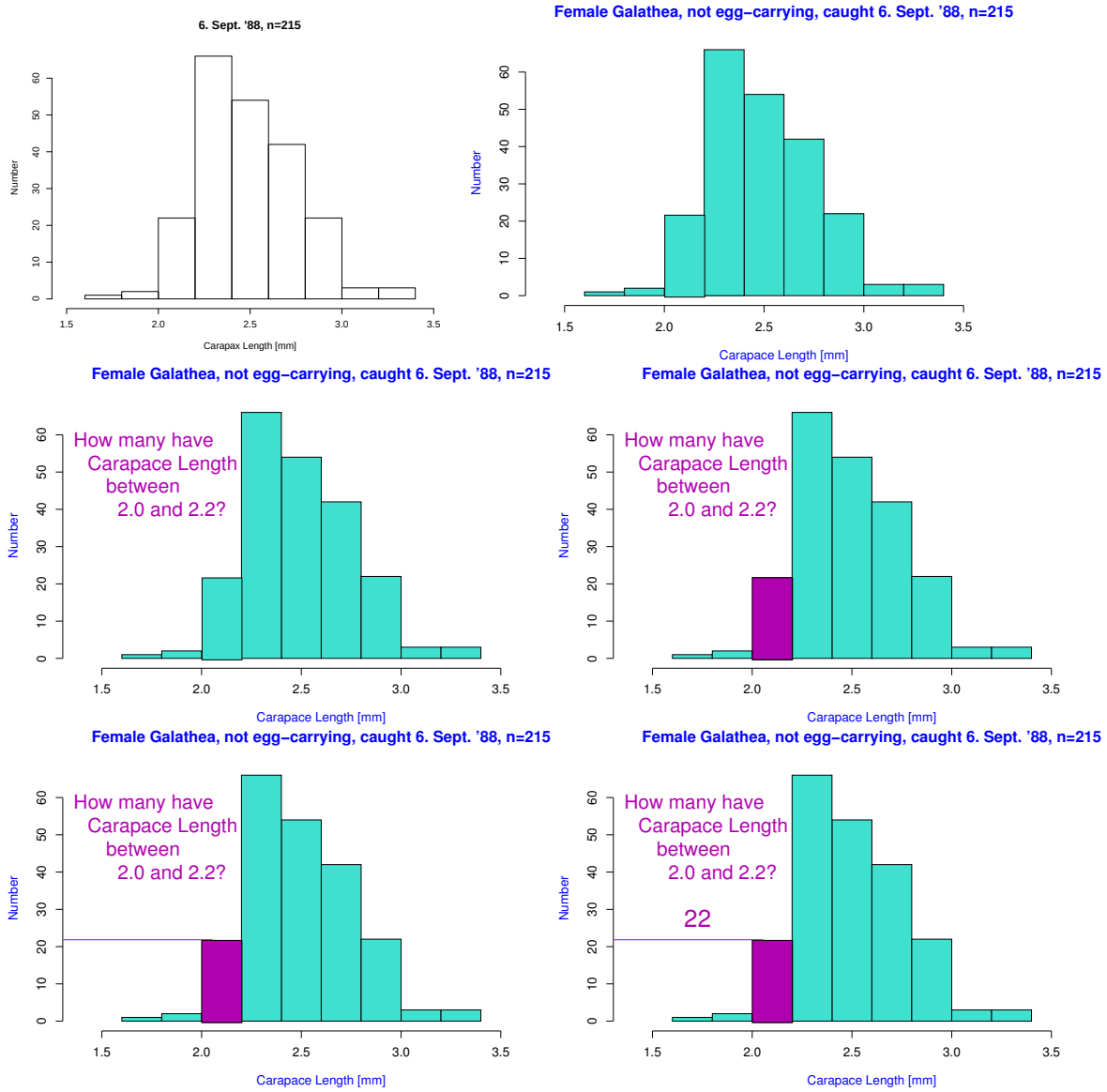
Helgoländer Tiefe Rinne, North Sea

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

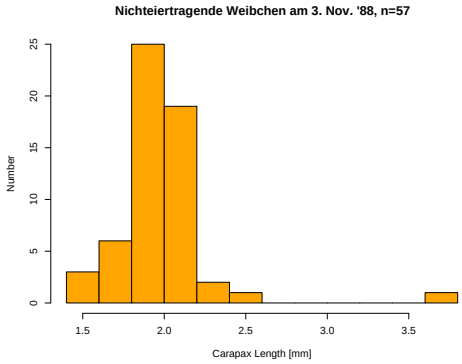
2.9	3.0	2.9	2.5	2.7	2.9	2.9	3.0
3.0	2.9	3.4	2.8	2.9	2.8	2.8	2.4
2.8	2.5	2.7	3.0	2.9	3.2	3.1	3.0
2.7	2.5	3.0	2.8	2.8	2.8	2.7	3.0
2.6	3.0	2.9	2.8	2.9	2.9	2.3	2.7
2.6	2.7	2.5	.	.	.	.	.



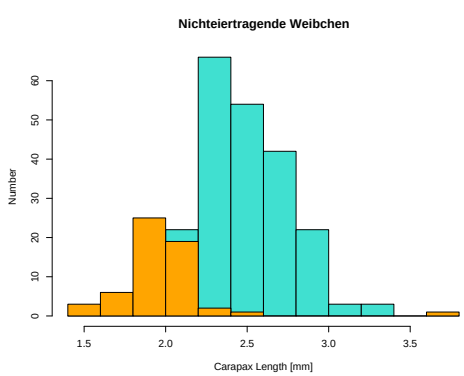
2.1 Histograms und Density Polygons

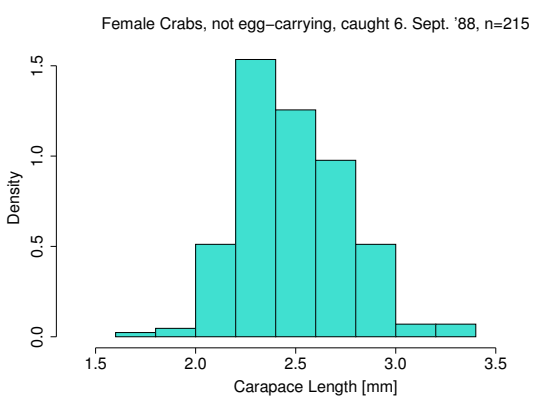


Two Months Later (3. Nov '88)

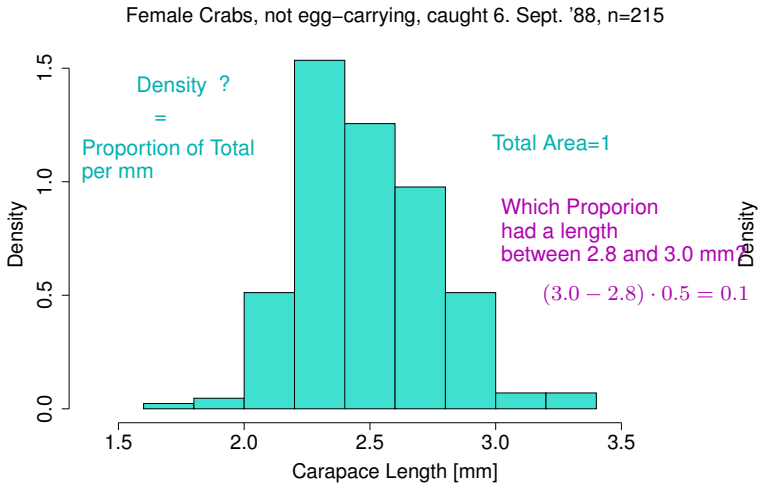


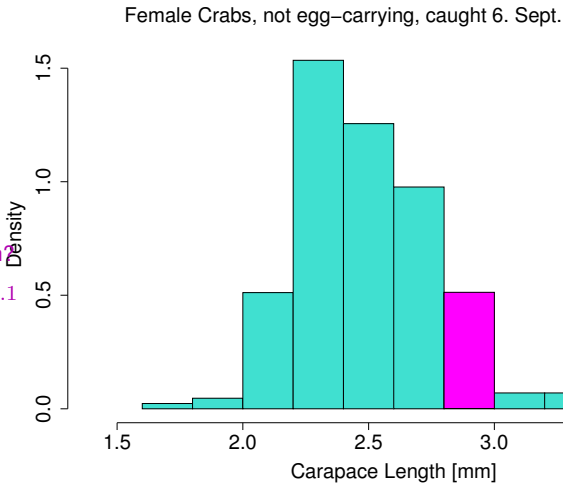
Comparing the two Distributions



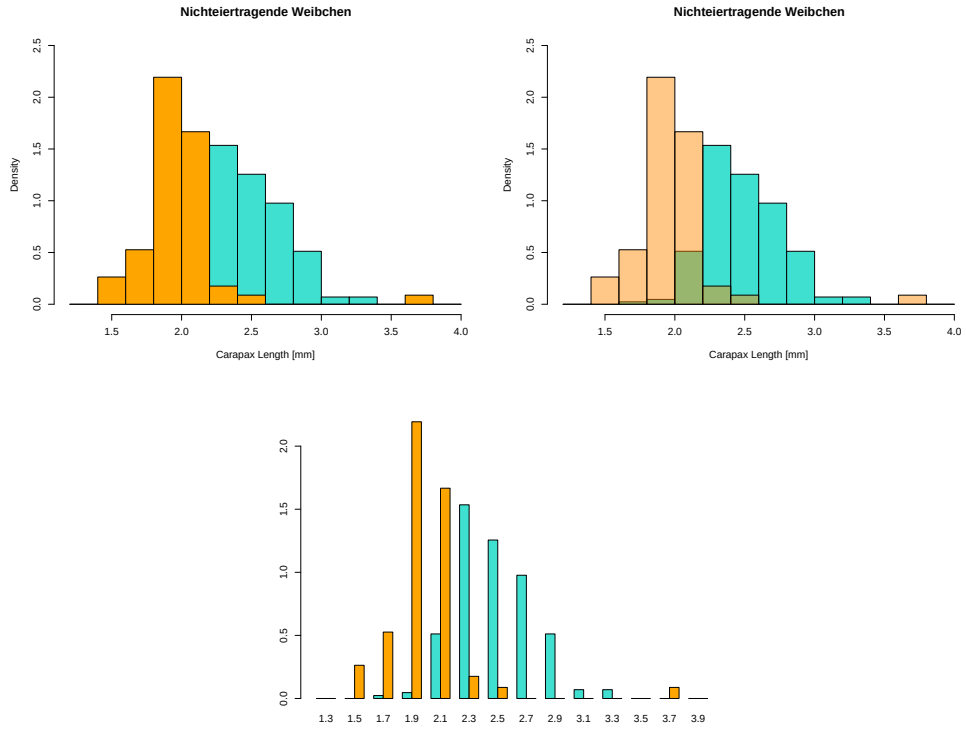


Problem: different sample sizes $\begin{matrix} 6.9.1988 & : & n = 215 \\ 3.11.1988 & : & n = 57 \end{matrix}$ Idea: scale y-axis such that each distribution has total area 1.


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]


Female Crabs, not egg−carrying, caught 6. Sept.
Density
0.0
0.5
1.0
1.5
1.5
2.0
2.5
3.0
Carapace Length [mm]

How to compare the two distributions?



My Advice

If you are a commercial artist:

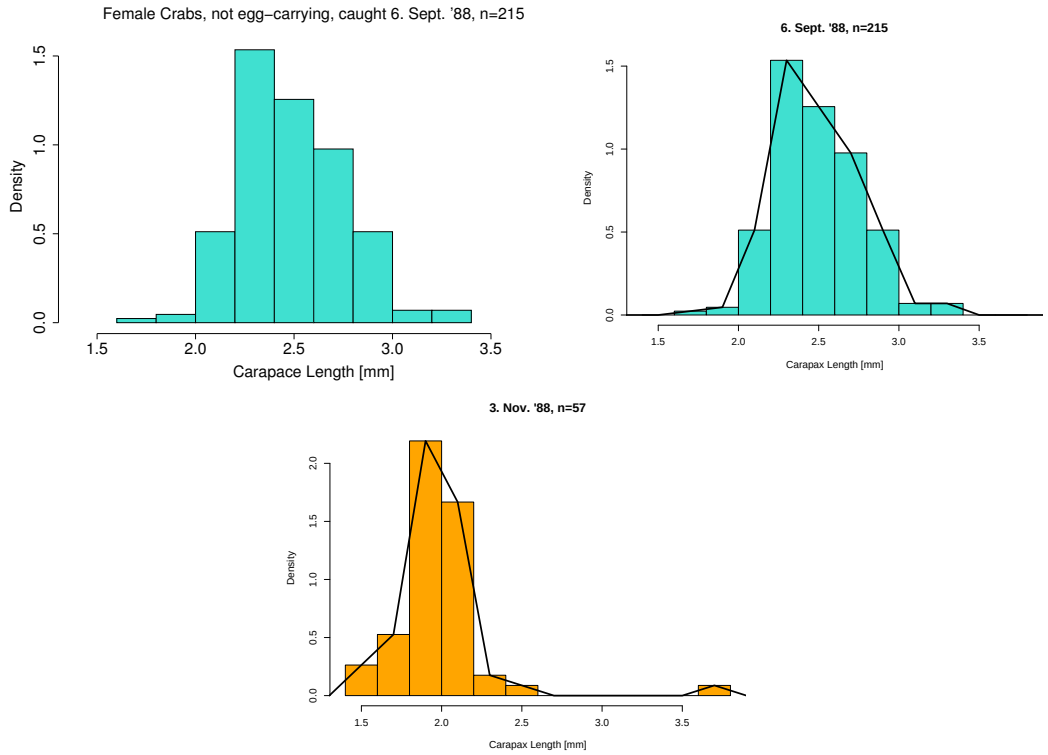
Impress everybody with cool 3D graphics!

If you are a scientist:

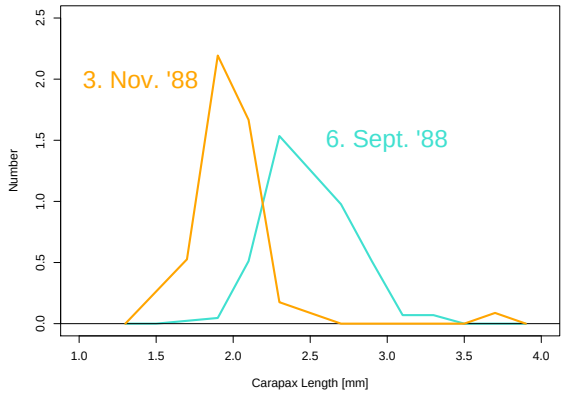
Visualize your data in clear and simple 2D plots.

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

Simple and Clear: Density Polygons



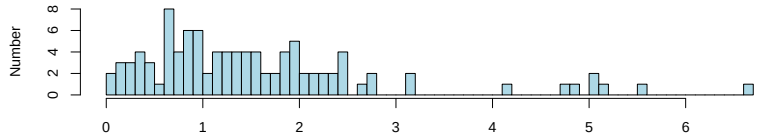
Convenient to show two or more Density Polygons in one plot

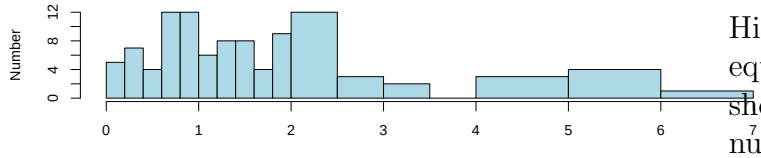


Biological Interpretation: What may be the reason for this shift?

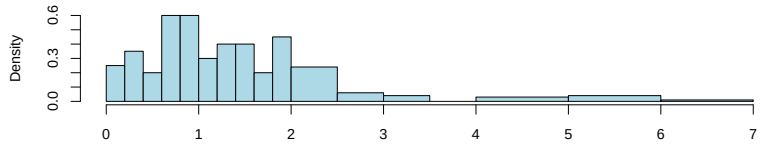
2.1.1 Histograms: Densities or Numbers?

Number vs. Density

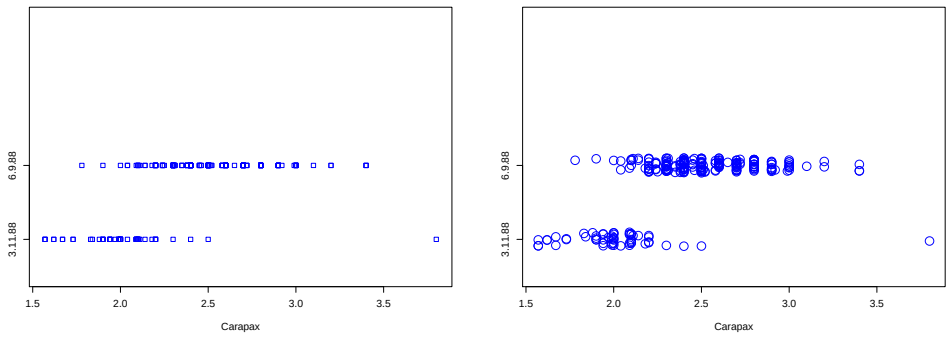


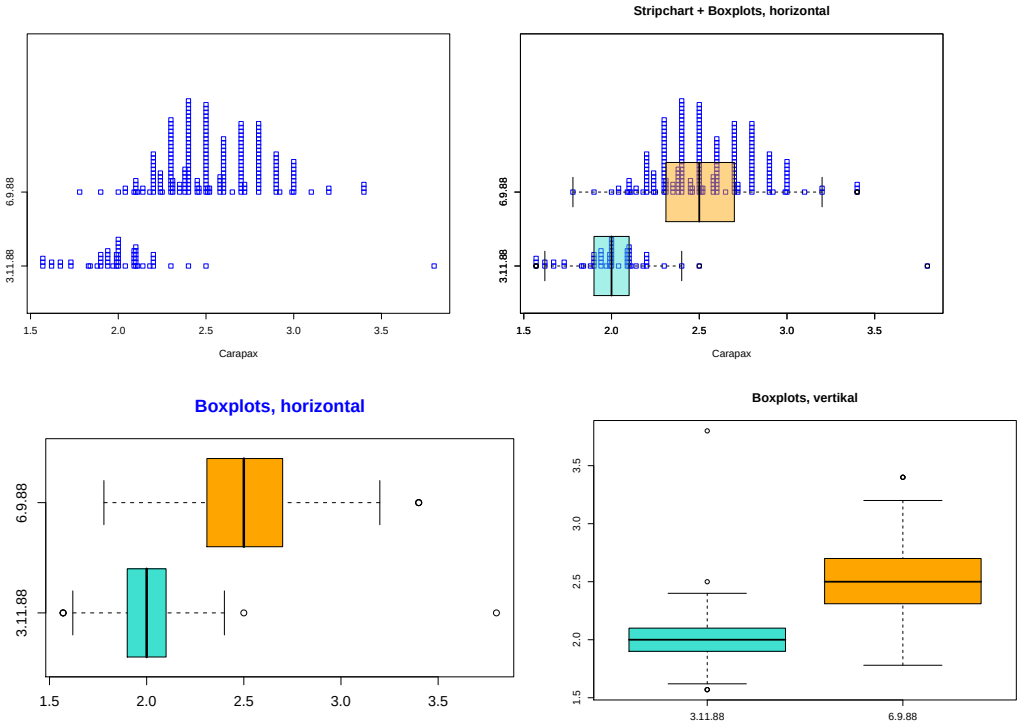


Histograms with unequal intervals should show densities, not numbers!



2.2 Stripcharts and Boxplots

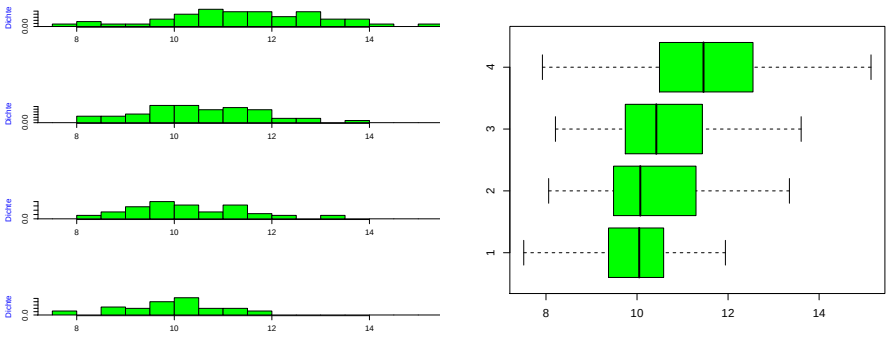




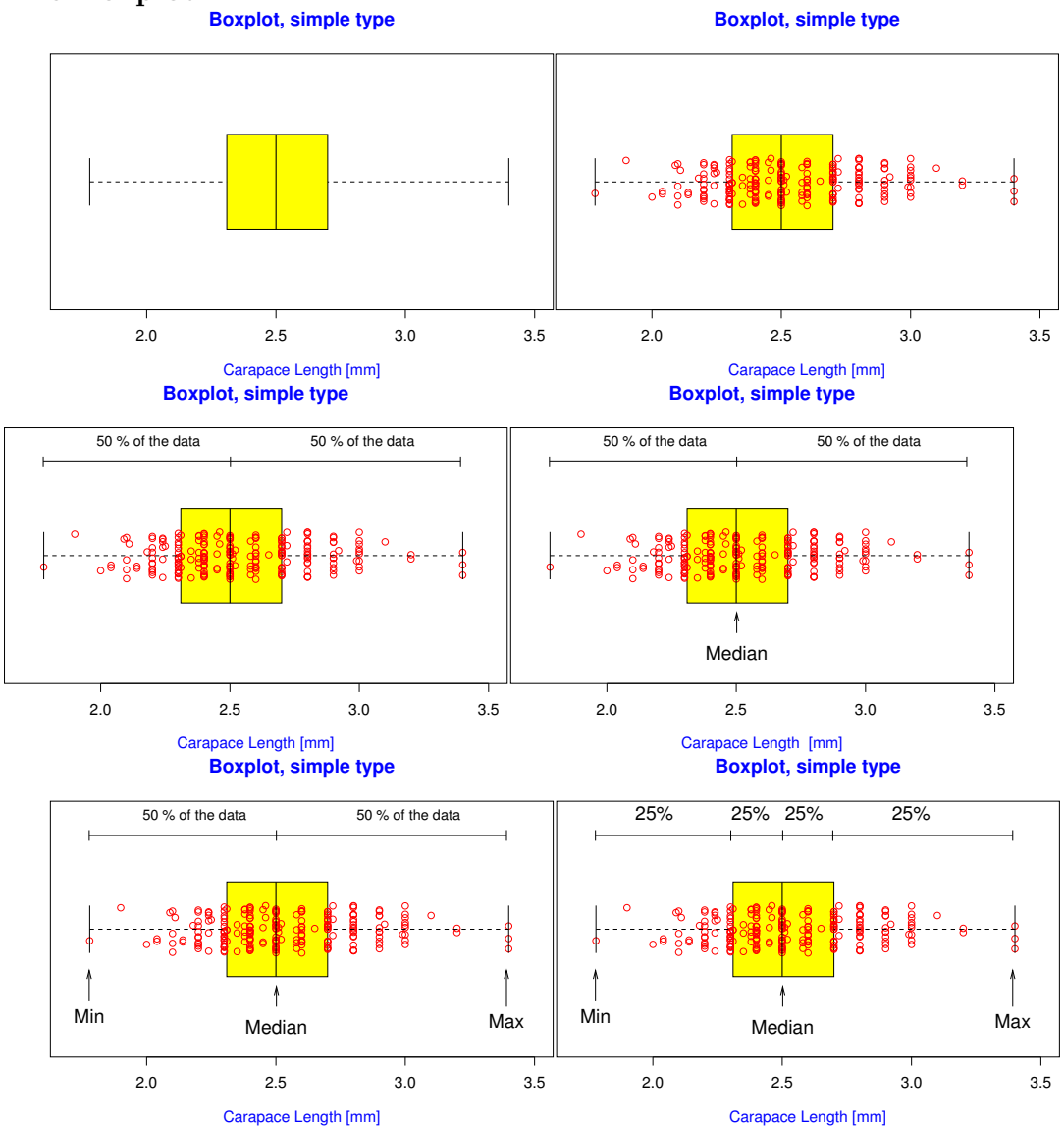
Simplify to understand

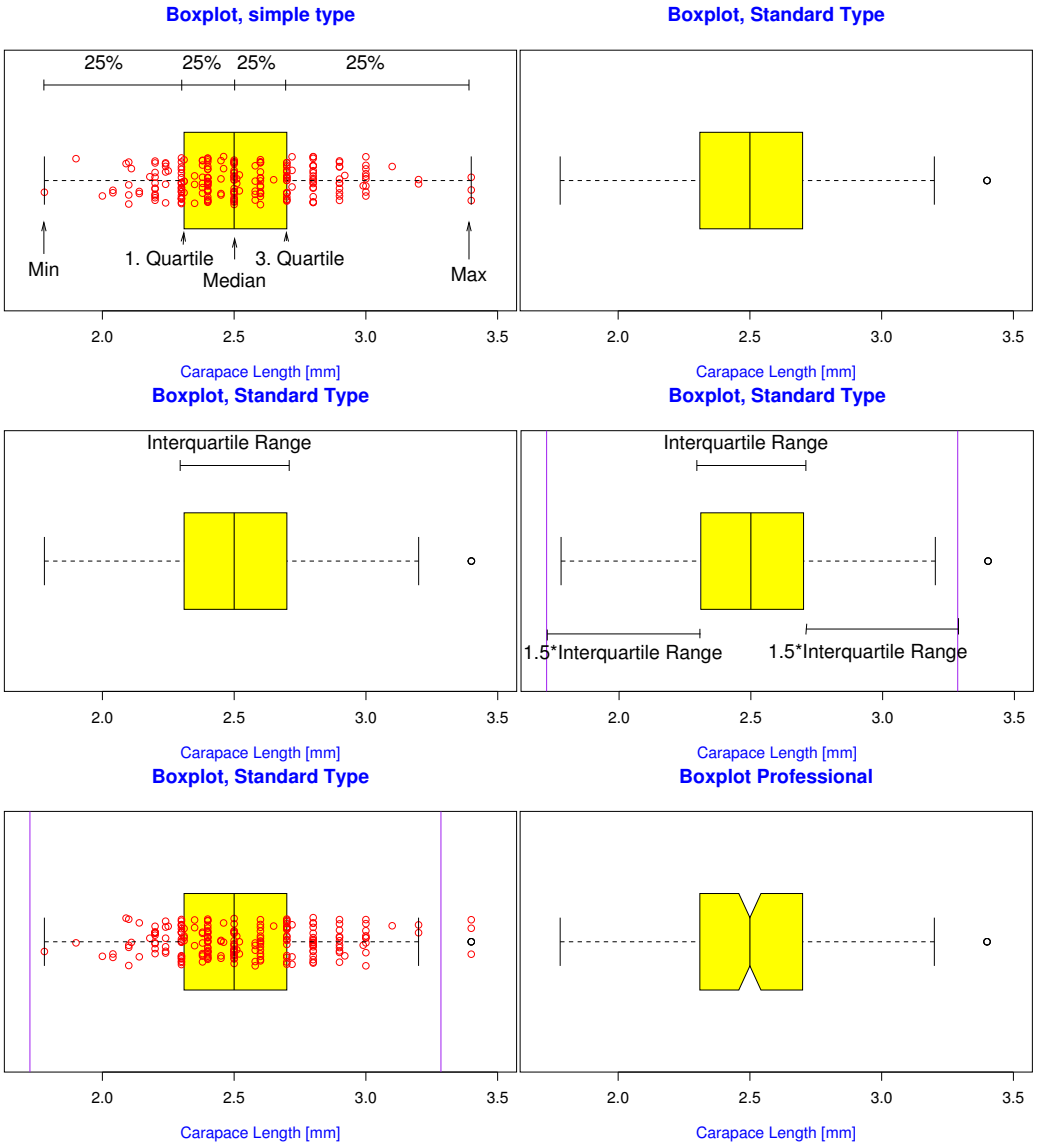
Histograms and density polygons
allow a comprehensive view on the data.
Sometimes too comprehensive.

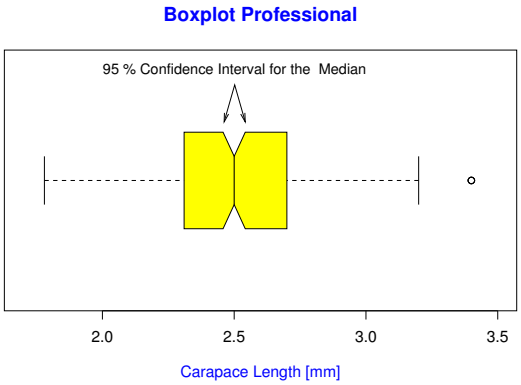
Comparison of four groups



The Boxplot


Boxplot, simple type
Boxplot, simple type
2.0
2.5
3.0
3.5
Carapace Length [mm]
Carapace Length [mm]
Boxplot, simple type
Boxplot, simple type
50 % of the data
50 % of the data
50 % of the data
50 % of the data
Median
Carapace Length [mm]
Carapace Length [mm]
Boxplot, simple type
Boxplot, simple type
50 % of the data
50 % of the data
25%
25%
25%
25%
Min
Median
Max
Min
Median
Max
Carapace Length [mm]
Carapace Length [mm]


Boxplot, simple type
25%
25%
25%
25%
Min
1. Quartile
Median
3. Quartile
Max
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, Standard Type
Interquartile Range
2.0
2.5
3.0
3.5
Carapace Length [mm]
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
2.0
2.5
3.0
3.5
Carapace Length [mm]



2.3 Example: Darwin Finches

Charles Robert Darwin (1809-1882)

Darwin Finches

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

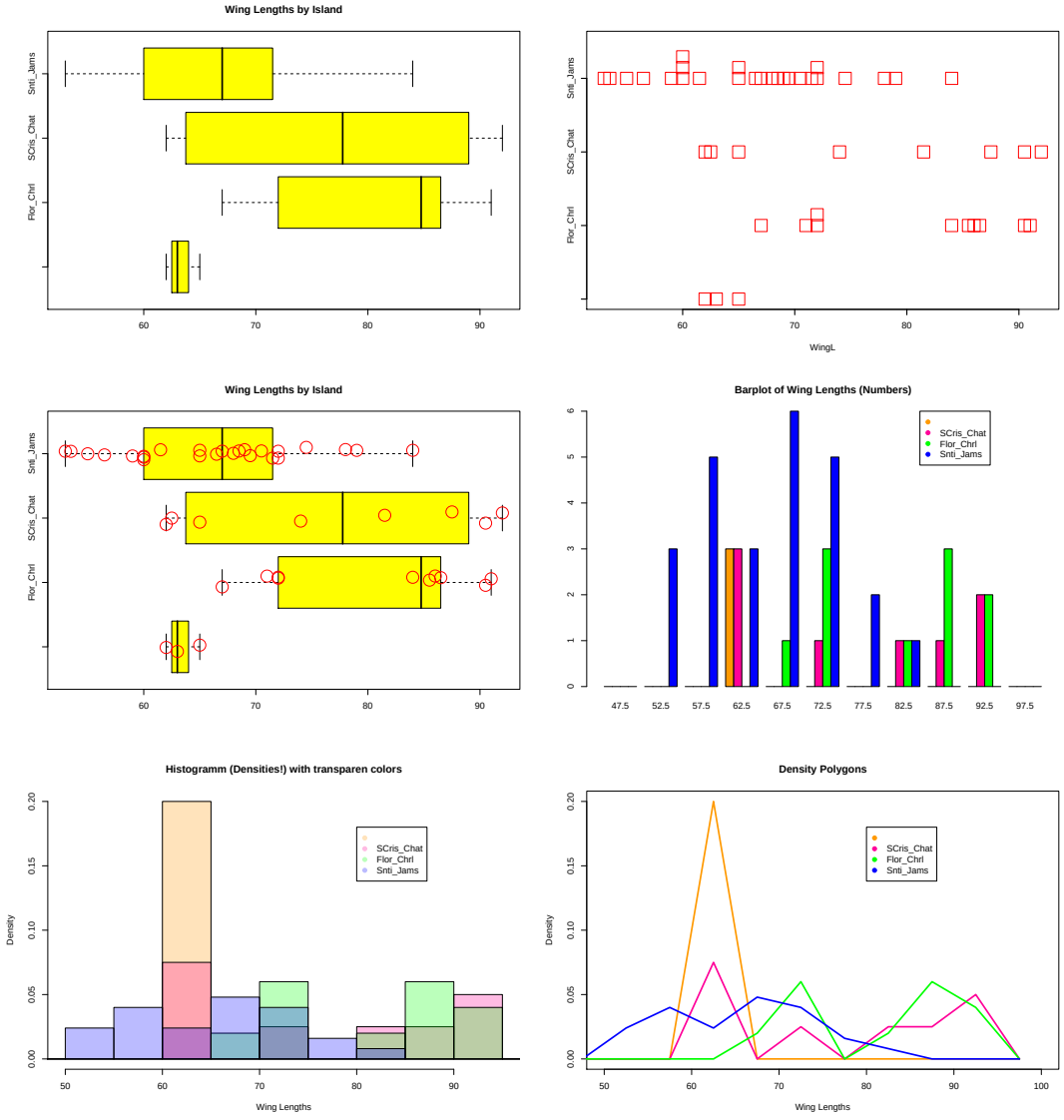
Darwin's collection of Finches

References

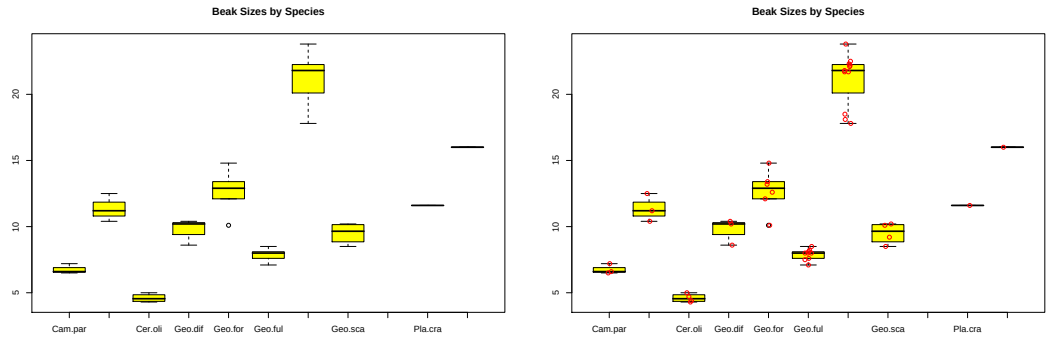
[1] Sulloway, F.J. (1982) The Beagle collections of Darwin's Finches (Geospizinae). *Bulletin of the British Museum (Natural History), Zoology series* **43**: 49-94.

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

Wing Sizes of Darwin's Finches



Beak Sizes of Darwin's Finches



2.4 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.

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:

Mean[1ex] =[1ex]

$$\frac{\text{Sum of oberved values}}{\text{Number of Observations}}$$

$$\frac{\text{Sum}}{\text{Number}}$$

The formula for the mean of $x_1, x_2, \ldots, x_n$:

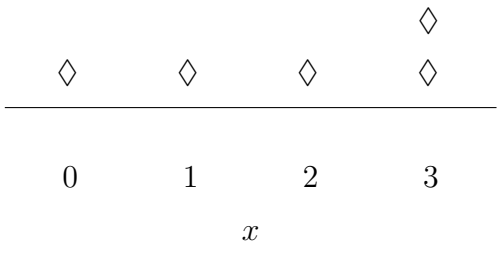
$$\begin{aligned} \overline{x} &= (x_1 + x_2 + \cdots + x_n)/n \\ &= \frac{1}{n}\sum_{i=1}^{n} x_i \end{aligned}$$

Geometric Interpretation of the Mean

Center of Gravity

Mean = Center of Gravity

Where is the center of gravity?



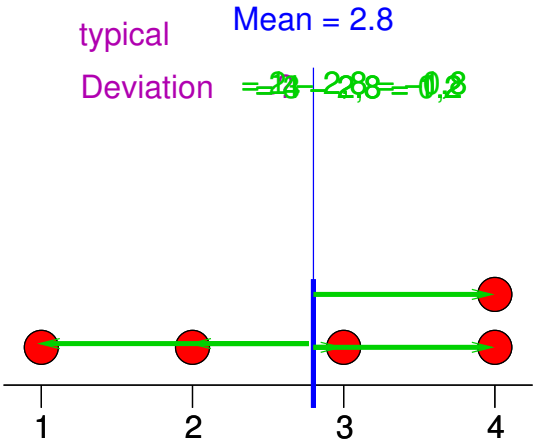
$m = 1.5$?

$m = 2$?

$m = 1.8$?

The Standard Deviation

How far do typical overservations deviate from the mean?



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

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

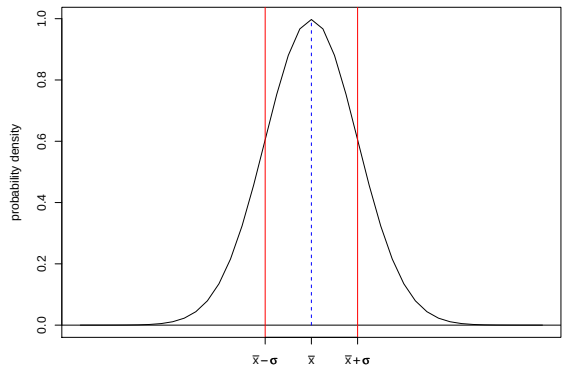
The formula for the *Standard Deviation* of $x_1, x_2, \ldots, x_n$:

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

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

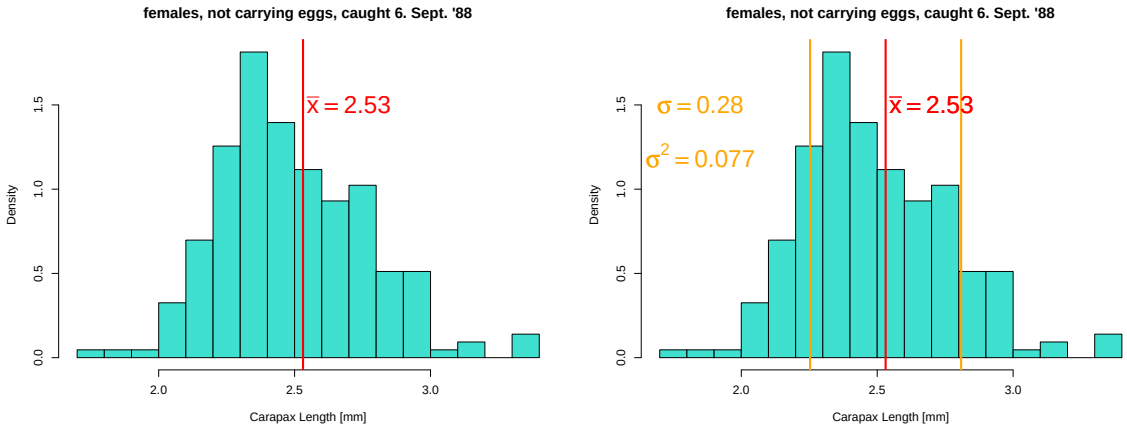
Rule of Thumb for the Standard Deviation

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 $\sigma^2_{\mathcal{X}}$, the variance in the whole population?Yes, we can! However, $\sigma^2_{\mathcal{S}}$ is on average by a factor of $\frac{n-1}{n}$ $(= 214/215 \approx 0,995)$ smaller than $\sigma^2_{\mathcal{X}}$.

Variances

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

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

(Corrected) Sample Variance:

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

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

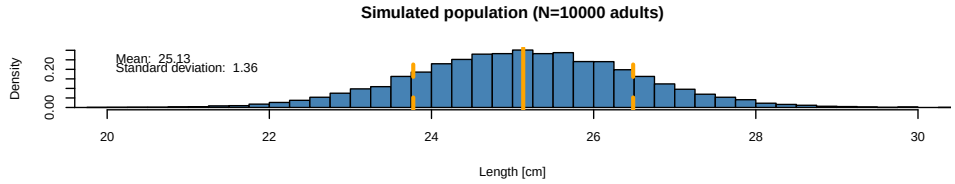
Example: Computing SD

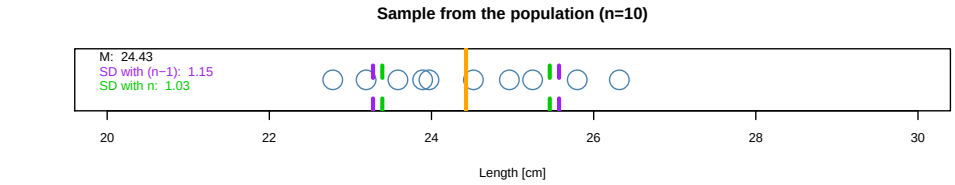
Given Data $\overline{x} =? \ \overline{x} = 10/5 = 2$

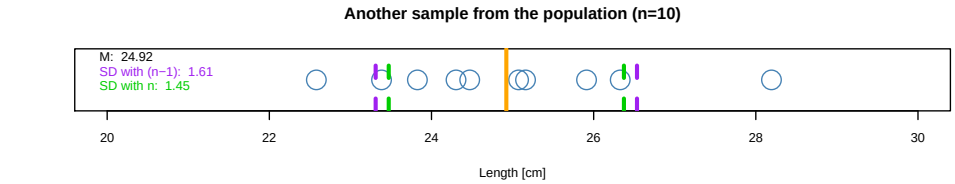
						$\sum$
x	1	3	0	5	1	10
$x - \overline{x}$	-1	1	-2	3	-1	0
$(x - \overline{x})^2$	1	1	4	9	1	16

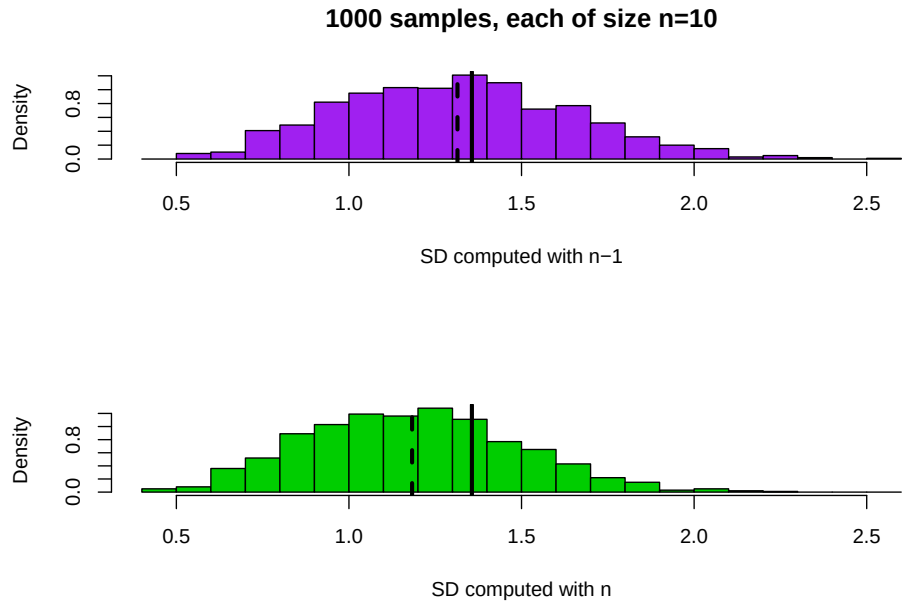
$$\begin{aligned} s^2 &= \sum \left((x - \overline{x})^2\right)/(n-1) \\ &= 16/(5-1) = 4 \\ s &= 2 \end{aligned}$$

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

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

3.3 Mean values are usually nice but sometimes mean

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.

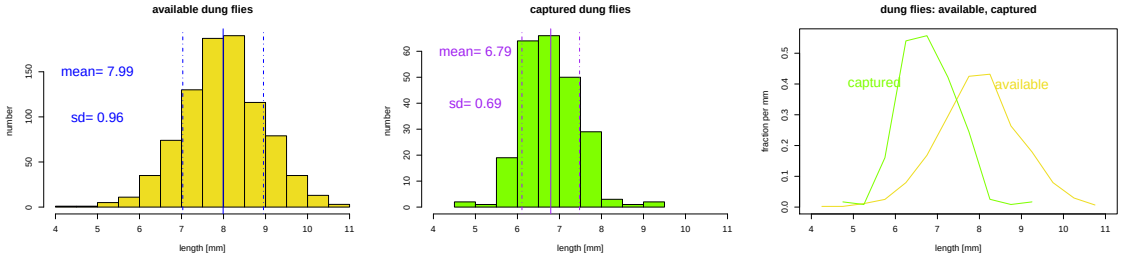
3.3.1 example: picky wagtails

Conjecture

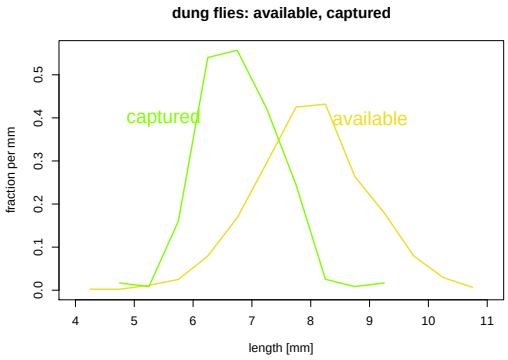
- Size of flies varies.
- efficiency for wagtail = energy gain / time to capture and eat
- lab experiments show that efficiency is maximal when flies have size 7mm

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

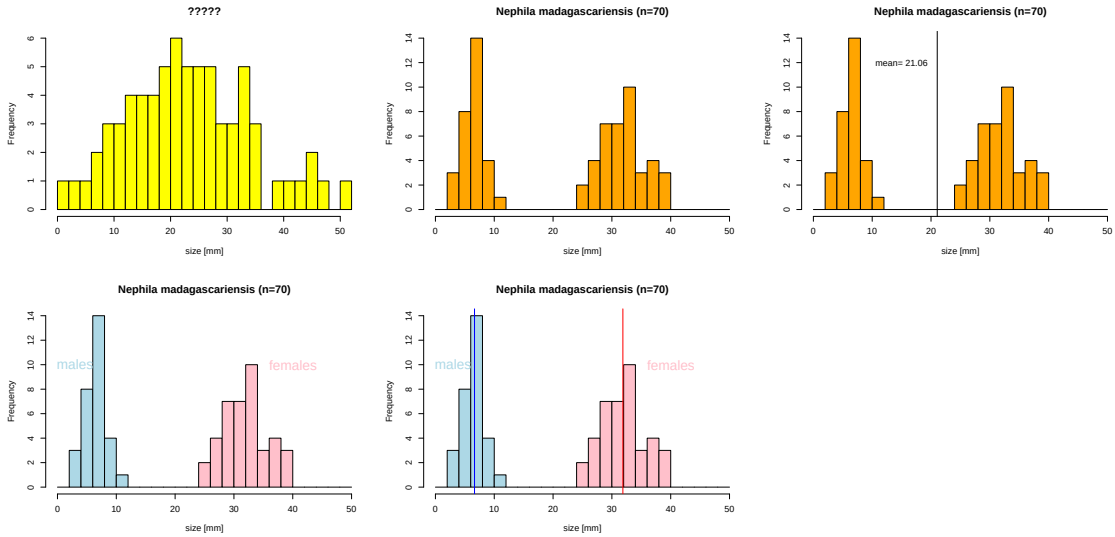
The birds prefer dung-flies from a relatively narrow range around the predicted optimum of 7mm.

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

3.3.2 example: spider men & spider women

Nephila madagascariensis

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.

3.3.3 example: copper-tolerant browntop bent

References

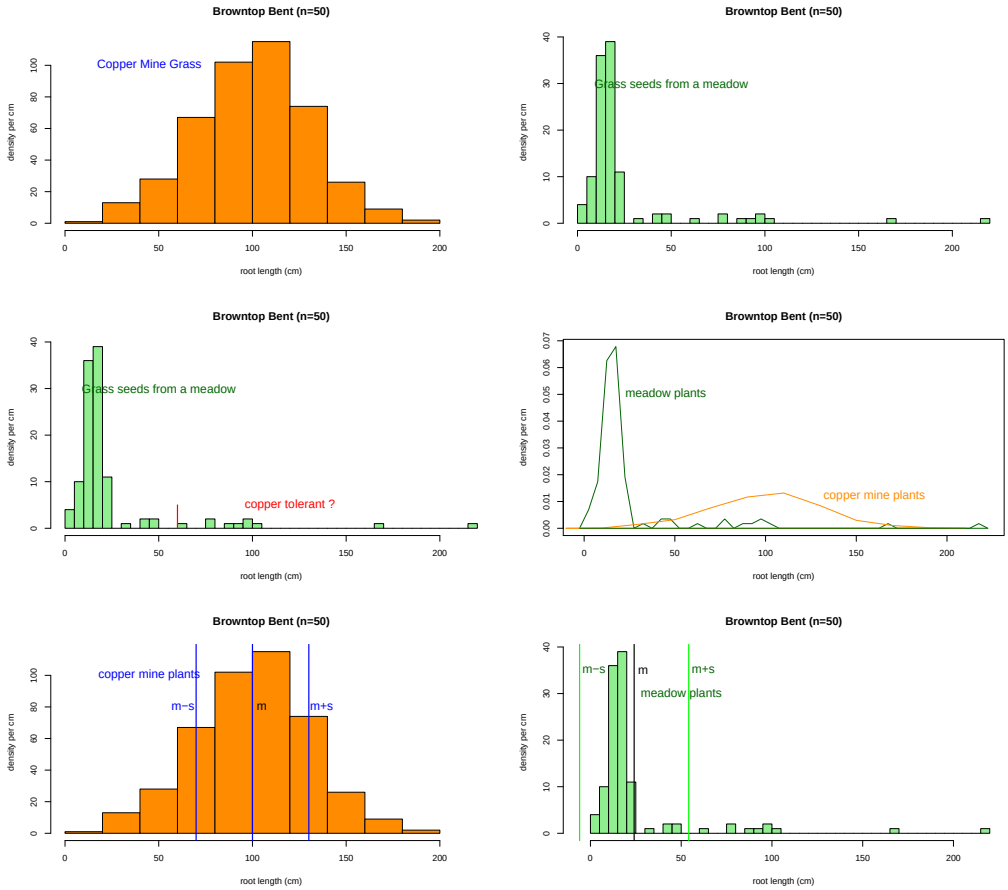
[Bra60] A.D. Bradshaw. Population Differentiation in *agrostis tenius Sibth.* III. populations in varied environments. *New Phytologist*, 59(1):92 – 103, 1960.

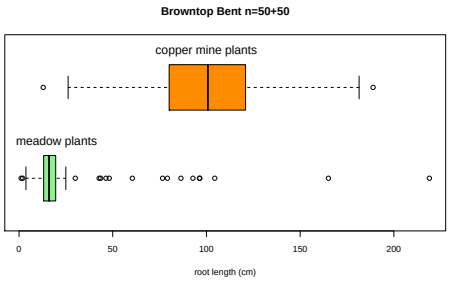
[MB68] T. McNeilly and A.D Bradshaw. Evolutionary Processes in Populations of Copper Tolerant Agrostis tenuis Sibth. *Evolution*, 22:108–118, 1968.

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





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 four quartiles
max, Q_1, median, Q_3, max
that are shown in the boxplot are more approriate.

Conclusions from this section

Always visually inspect the data!

Never rely on summarising values alone!