Data analysis with R

Lecture 4

Introduction to Data Analysis

Jouni Junnila

Exploratory Data Analysis

- The first step in data analysis is Exploratory Data Analysis (EDA).
- EDA is a name for a collection of data display techniques that are intended to let the data speak for themselves, prior to, or as a part of formal analysis.
- EDA looks for what may be apparent form a direc t, careful and assumption-free examination of the data.

Four roles of EDA

1) EDA examines the data. This examination might suggest how data should be analyzed or interpreted. This fits well with the view of science as inductive reasoning.
2) EDA results might challenge the theoretical understanding that guided the data collection. In this case EDA has a more revolutionary role.
3) EDA allows the data to influence and criticize the intended analysis. EDA facilitates checks on assumptions: formal analysis can then proceed with confidence.
    - Note, that in some fields (eg. Clinical trials) the analysis methods must be decided beforehand. This limites the meaning of EDA

Four roles of EDA

4) EDA might reveal additional information about the data, not directly related to the research question.
   - May suggest new lines of research

- However a data analyst must keep some sense in their EDA i.e. there is a risk of looking too hard. We practically can always see some patterns in the data, if we ”torture” the data enough.
- EDA must therefore remain as a first step of data analysis, not the main part.

Viewing a single dataset

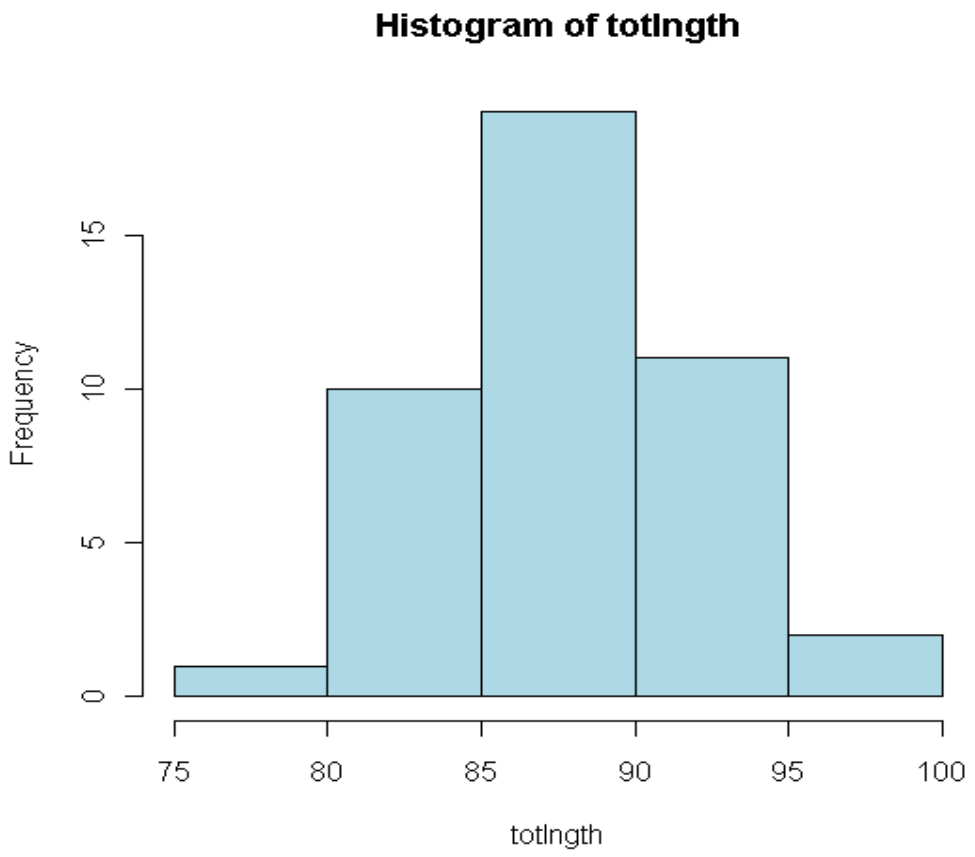
- When we first get a new dataset in our hands, the first thing to do is to start drawing some graphics from the data.
- Next we'll go through some of the basic graphical methods that statisticians most commonly use to familiarize themselves with the current data at hand.

Histograms

- Histograms are one of the very basic EDA-tools.
- Gives a graphical representation of the frequency distribution of the data.
- A bell-shaped histogram suggests that the data is from the normal distribution.
- Not so easy to interpret with small samples.
- Viewer can be easily fooled by selection of break points. It’s a good practice to always test a few different break points.

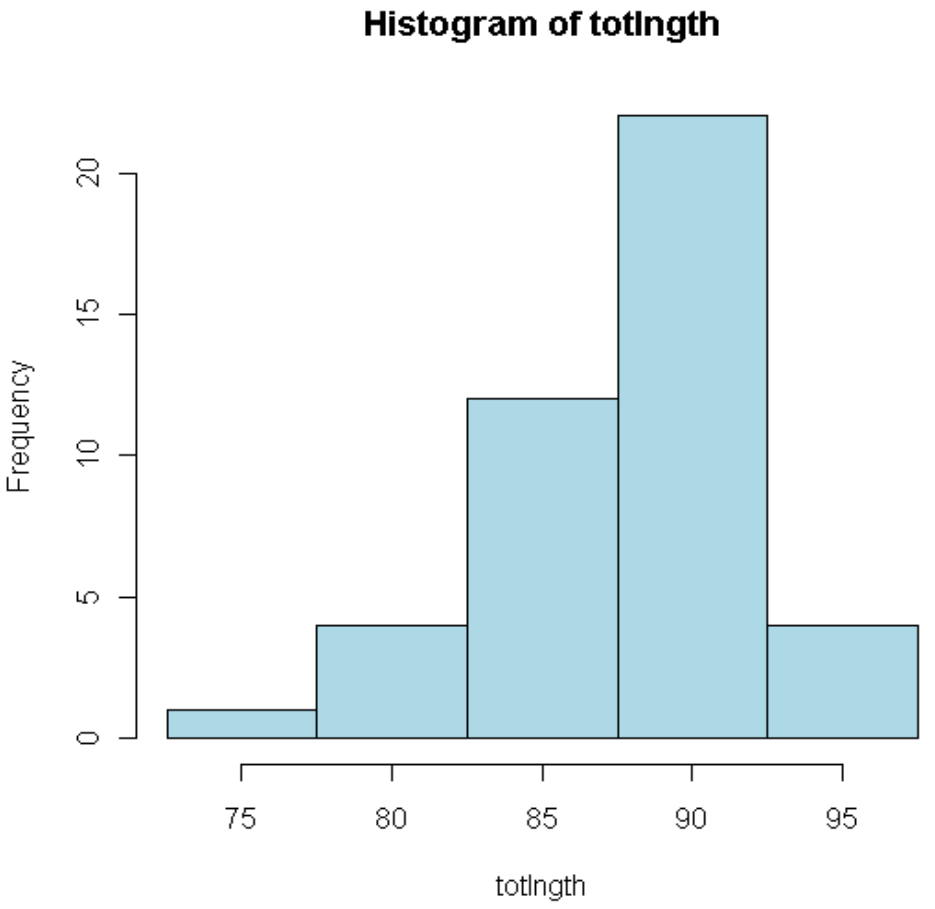
Histogram; example (a)

- Histogram of total length with break-points 75, 80, 85, etc.
- The data seems to be normally distributed
- `hist(totlngth, col="lightblue")`



Histogram; example (b)

- Histogram of same data, but with break-points 72.5, 77.5, etc.
- Looks quite different than the previous histogram!
- `hist(totlngth, col="lightblue", breaks=72.5 + (0:5) * 5)`



Density plots

- Histograms are crude estimates of density.
- To get a better estimate we can draw a smooth density estimate.
- With histograms we have to choose the width of histogram bars somewhat subjectively, whereas density estimates require a bandwidth parameter that controls the amount of smoothing.
  - R-defaults works usually quite well for bandwidth parameter.

Histograms & densities

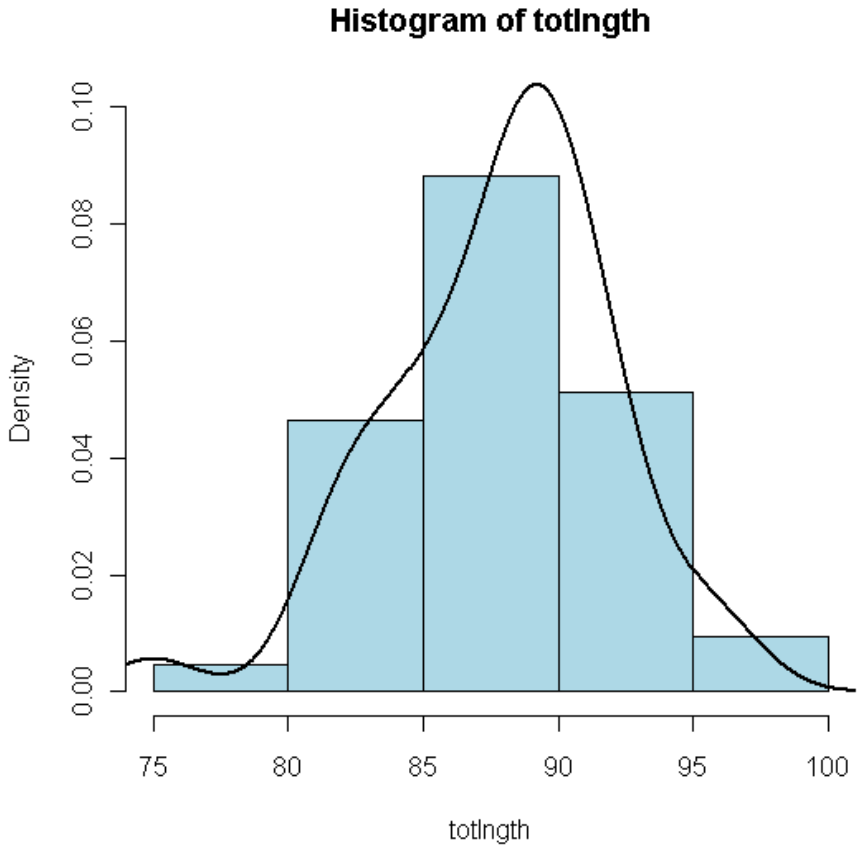
- After calculating the density estimates we can overlay them with the previously drawn histograms.
- With these kind of graphs we can draw attention to certain forms of non-normality,
  - i.e skewness or kurtosis in the data.
- Non-normality is a problem with many different statistical methods.
  - Therefore checking for normality is important!

Histograms & densities (2)

- Deciding whether the histogram is close enough to a normal distribution can be difficult, and is based only on the eye of the viewer.
  - Thus, the correctness of the decision is not absolutely sure.
- Therefore, there are several formal statistical tests for testing normality of the data as well. These will be covered later on.

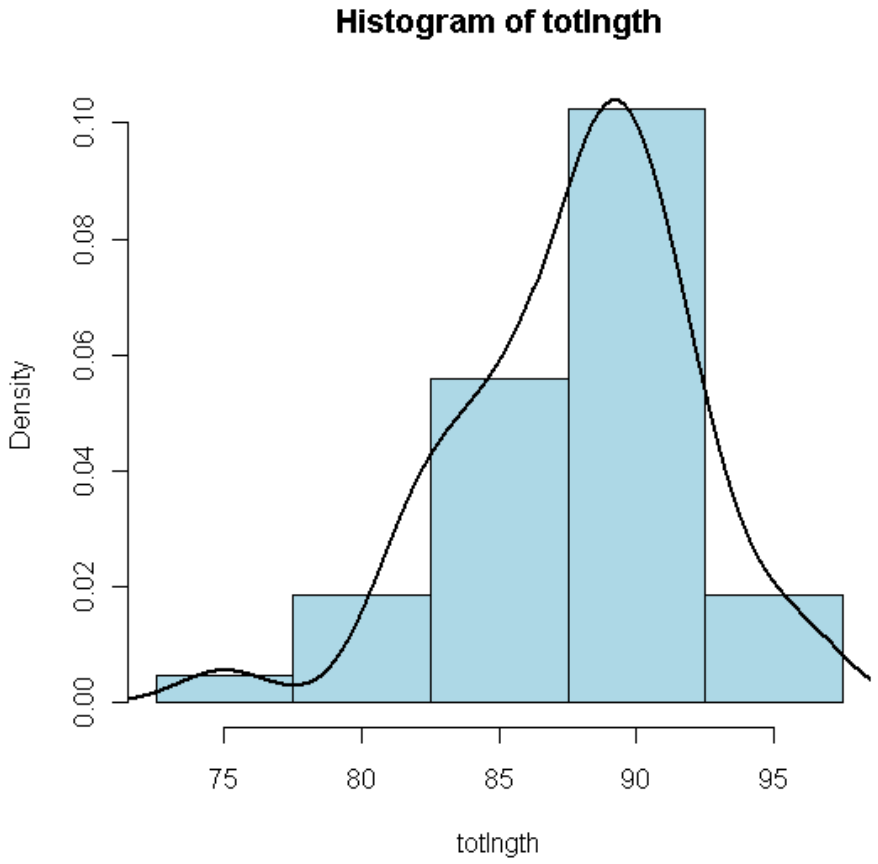
Histograms & densities; example (a)

```
dens <-
density(totlngth)
hist(totlngth,
col=”lightblue”,
freq=F)
lines(dens)
```



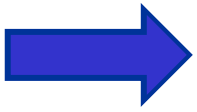
Histograms & densities; example (b)

```
hist(totlngth,
col="lightblue",
breaks=72.5 +
(0:5)*5, freq=F)
lines(dens)
```



Steam-and-leaf display

- The stem-and-leaf display is an alternative for histogram.
- It is used for displaying a single column of numbers.
- The resulting graph can be quite informative, but often quite hard to interpret, especially for non-statisticians



Therefore it is quite rarely used in actual data analysis

Steam-and-leaf display; example

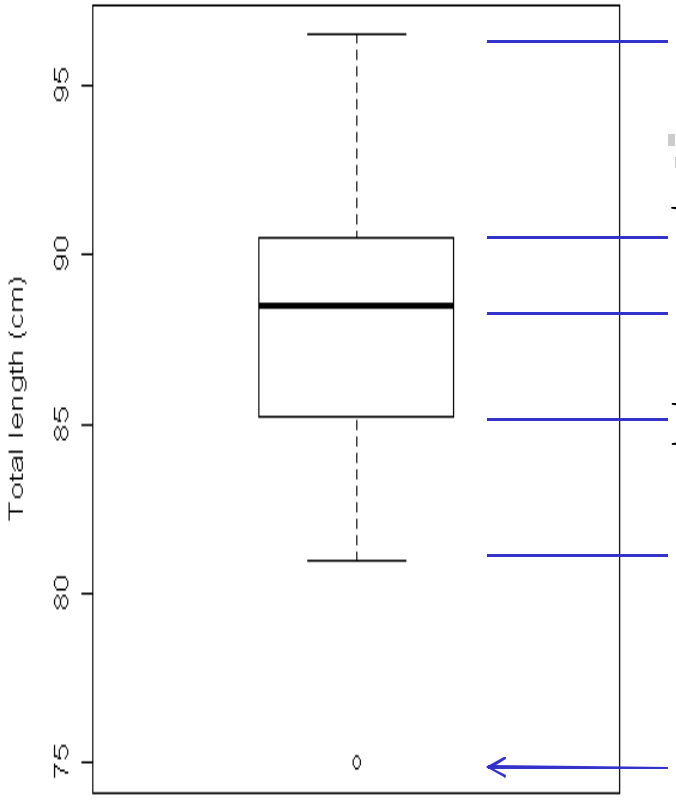
```
The decimal point is
at the |
74 | 0
76 |
78 |
80 | 05
82 | 0500
84 | 05005
86 | 05505
88 | 0005500005555
90 | 5550055
92 | 000
94 | 05
96 | 5
```

- stem(totlngth, scale=2)
- With the scale-parameter the length of the graph.

Boxplots

- Boxplots are a great way of representing data.
- It gives us a rough summary of the data.
  - With a trained eye you can comprehend the data at a short glance.
- Boxplots focus on specific important features of the data.
- It can also be used when we want to compare distributions of some variable in some subgroups (eg. female/male)

Boxplot; annotation



Smallest value (outliers excepted)

Upper quartile

Median

Lower quartile

Smallest value (outliers excepted)

Outlier

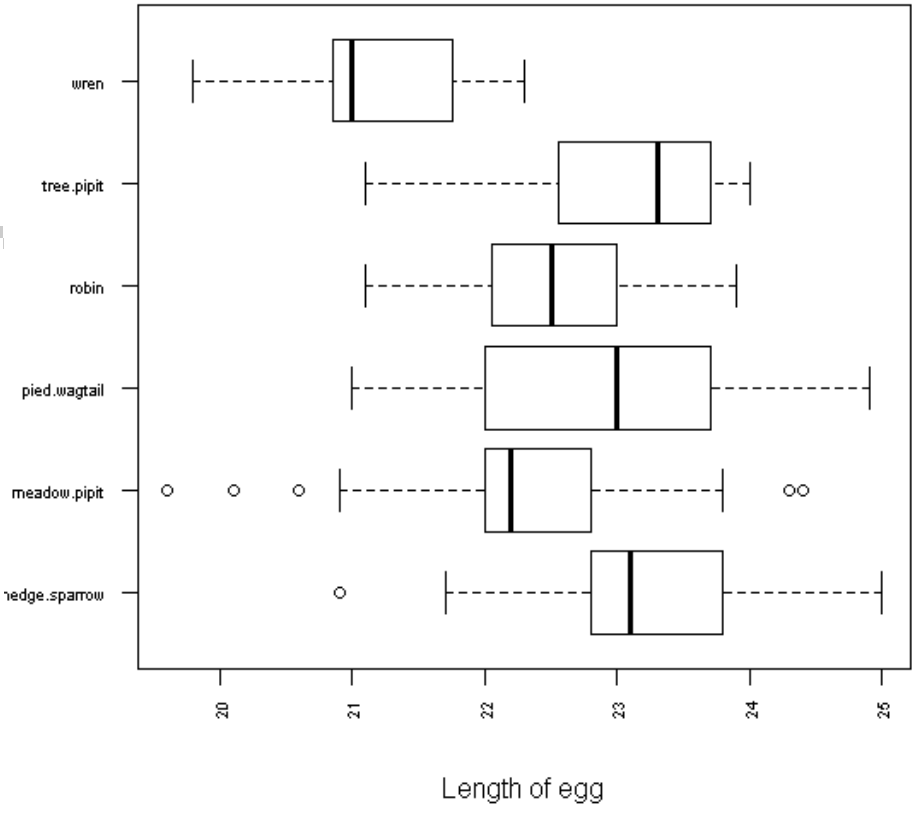
Boxplots; calculation

- Boxplots are also called box and whisker plots.
- Whiskers are the lines outside the box.
  - Maximum length of the whiskers is 0.75*interquartile range
  - The actual length = maximum/minimum value from the data that is smaller/bigger than the maximum length of the whisker.
  - Values bigger than the maximum length are considered outliers which are drawn separately (circles in R by default).

Boxplots; example

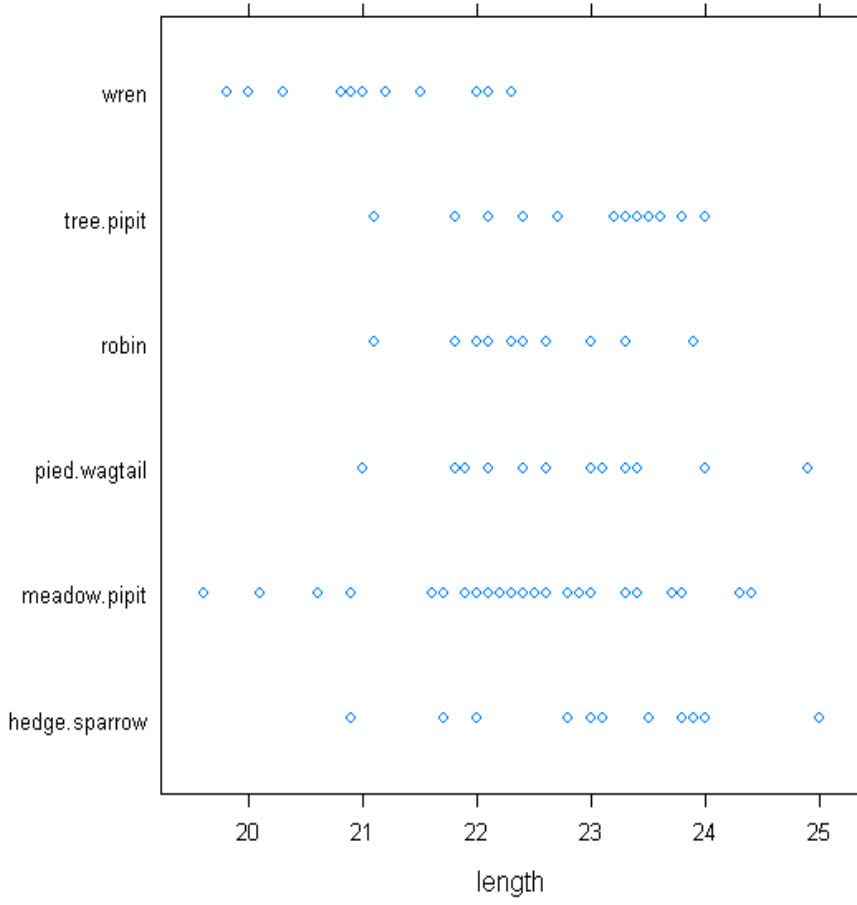
- We can compare groups determined by a factor in some numerical variable, e.g.

```
boxplot(length~species
data=cuckoos,
xlab="Length of egg",
horizontal=T,las=2)
```



stripplot()

- Function stripplot from the lattice-package gives the same kind of graph as boxplot, but doesn't summarize anything.
- All the observations will be drawn. (in the same way as outliers in boxplot)



Pairwise relationships

- In science we often want to examine the pattern and relationship of two numeric variables.
- A simple but important tool for that is scatter plot.
- With scatter plots it is important to select the right scale.
  - With wrong selection of scale, we can either show some pattern in the data, that isn't there or hide some pattern that there actually is.

Scatter plots

- Simple example of the use of scatter plot.
- `plot(length~breadth, data=cuckoos)`
- Note! If we would use species instead of breadth, we would get a boxplot. The type of graph is determined by the class of the variable.