Data Exploration using R

Statistics Refresher Workshop

Kai Xiong

k.xiong@auckland.ac.nz
Statistical Consulting Service
The Department of Statistics
The University of Auckland

July 1, 2011

Exploring your data

- Checking the data for patterns, relationships, structures, and other features.
- Can be done through graphs and summary statistics.
- We recommend R for generating graphs, much more flexible than Excel, SPSS, SAS.

Roles of Exploratory Graphics

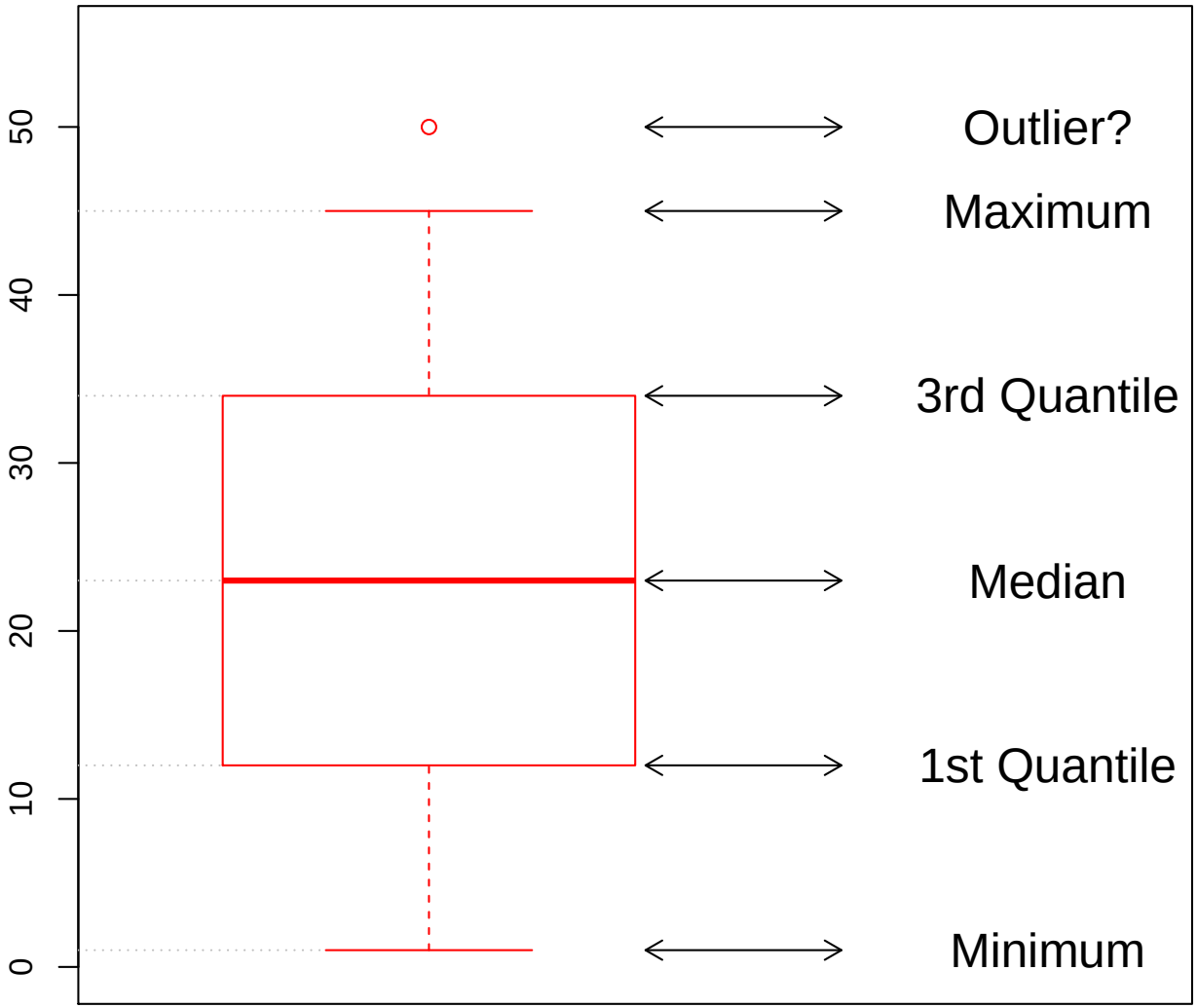
1. Are there any errors or outliers?
2. Are there any patterns in the data?
   - Symmetric, Skewed, Bimodal, Clusters?
3. Are there any relationships between the variables?
   - Linear (increasing, decreasing), polynomial, exponential...
4. What sort of model might be appropriate?

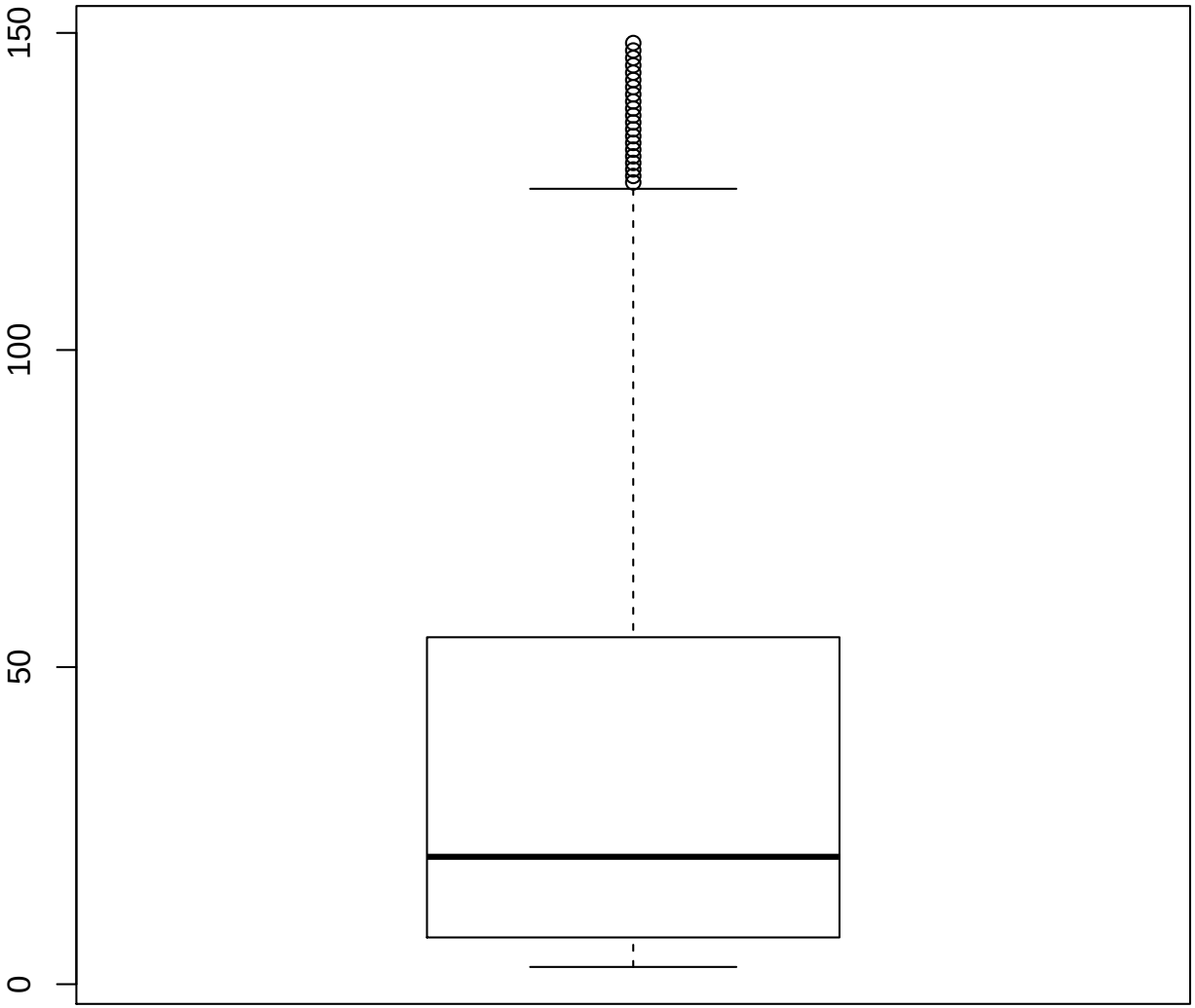
Types of variable

1. Quantitative
   - Continuous:
     e.g. gene expression, Body length of mussels
   - Discrete:
     e.g. number of photophores in lantern fish
2. Qualitative
   - Categorical
     e.g. SNPs, Location of marine reserves
   - Ordinal
     e.g. Position in a food chain

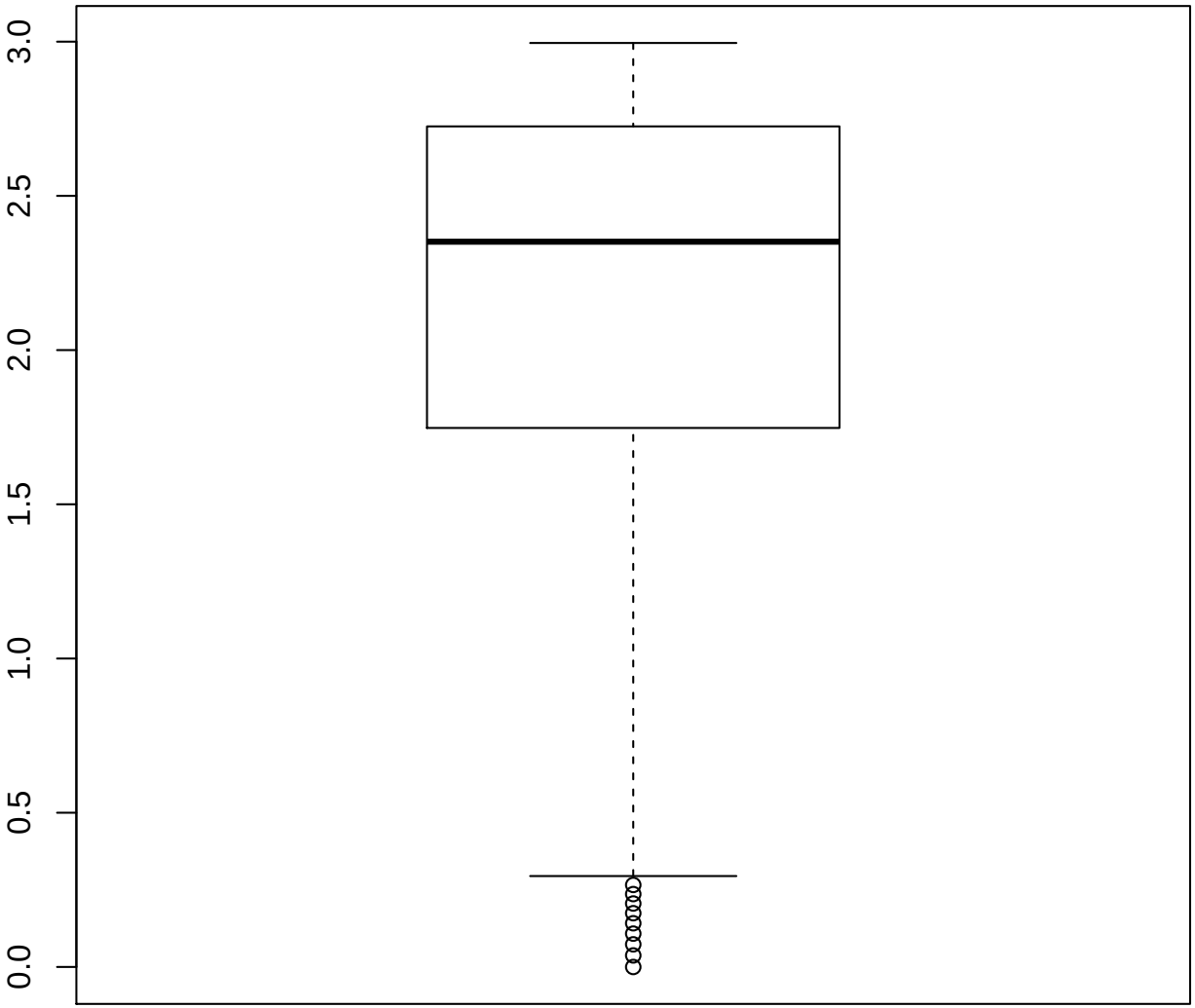
One Quantitative Variable

- Dotplots
- Five-number-summary statistics
  1. Minimum
  2. 1st Quantile/Lower Quantile/25th percentile
     - 25% of the sorted variable is smaller than the 1st quantile
     - 75% of the sorted variable is greater than the 1st quantile
  3. 2nd Quantile/Median/50th percentile
     - Divide the sorted variable in two equal halves
  4. 3rd Quantile/Upper Quantile/75th percentile
     - 25% of the sorted variable is greater than the 3rd quantile
     - 75% of the sorted variable is smaller than the 3rd quantile
  5. Maximum
- Boxplot: visual display of 5-number-summary statistics.
- Normal QQ plot
- Histogram

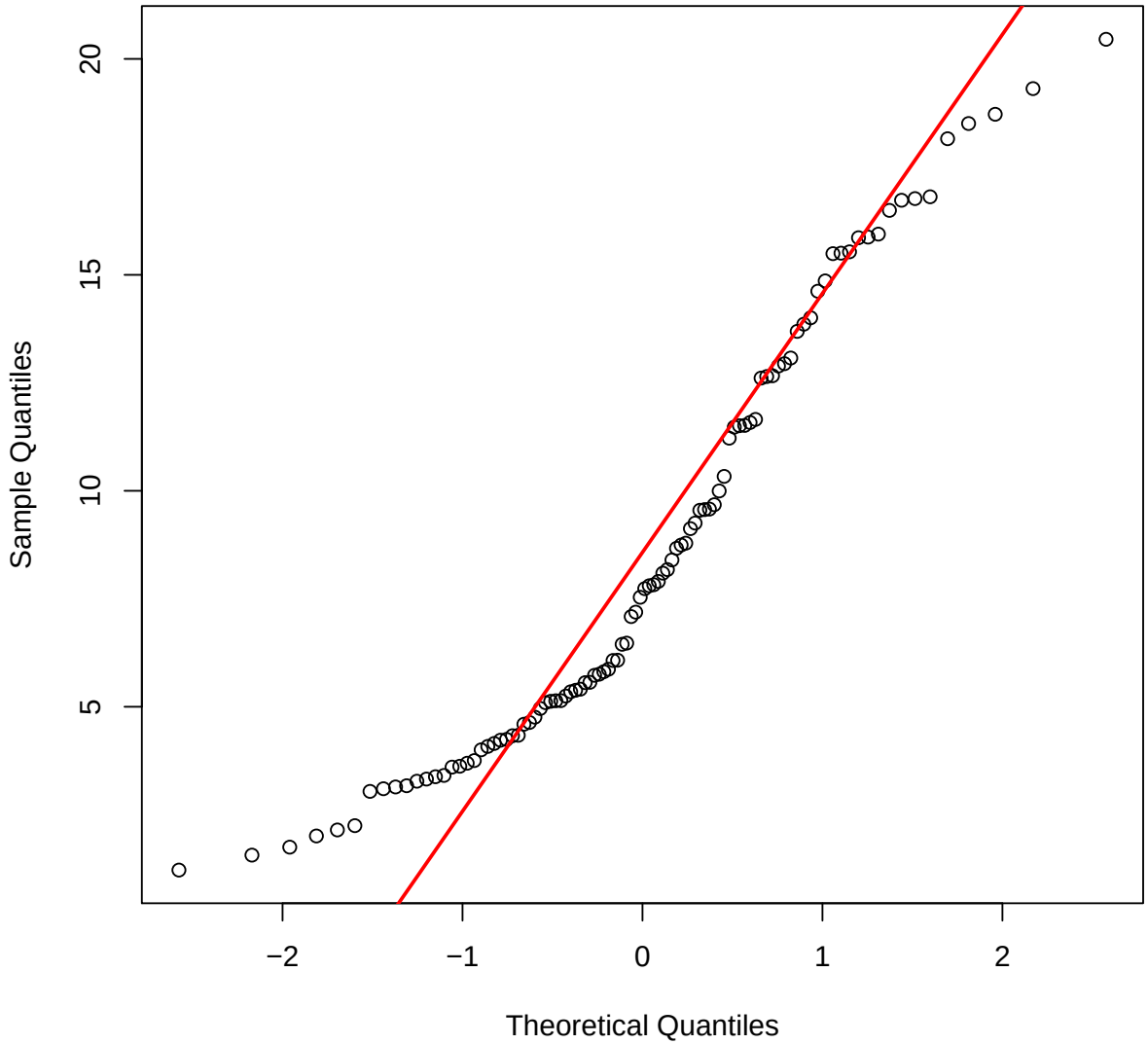

0
10
20
30
40
50
Outlier?
Maximum
3rd Quantile
Median
1st Quantile
Minimum

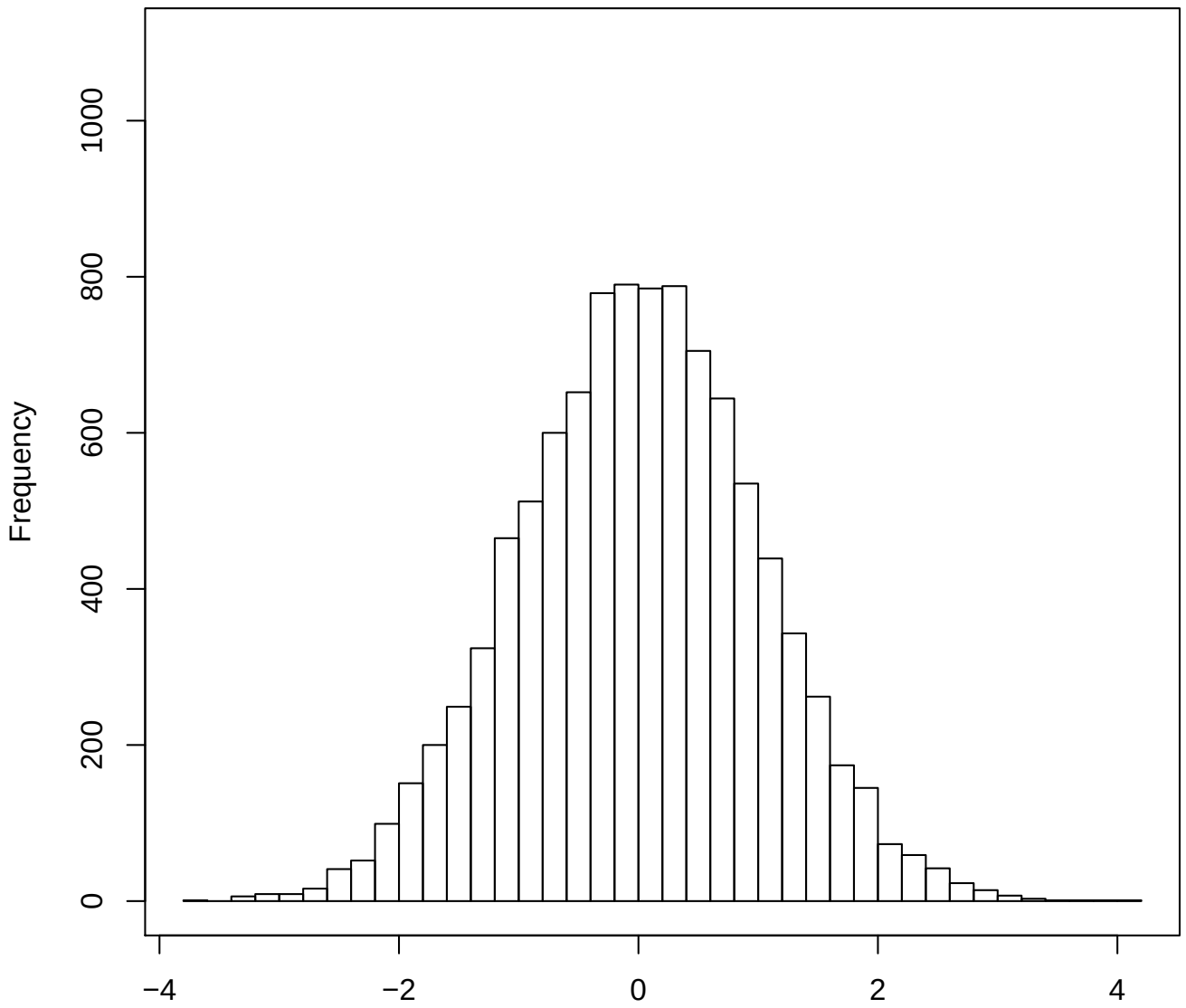

150
100
50
0

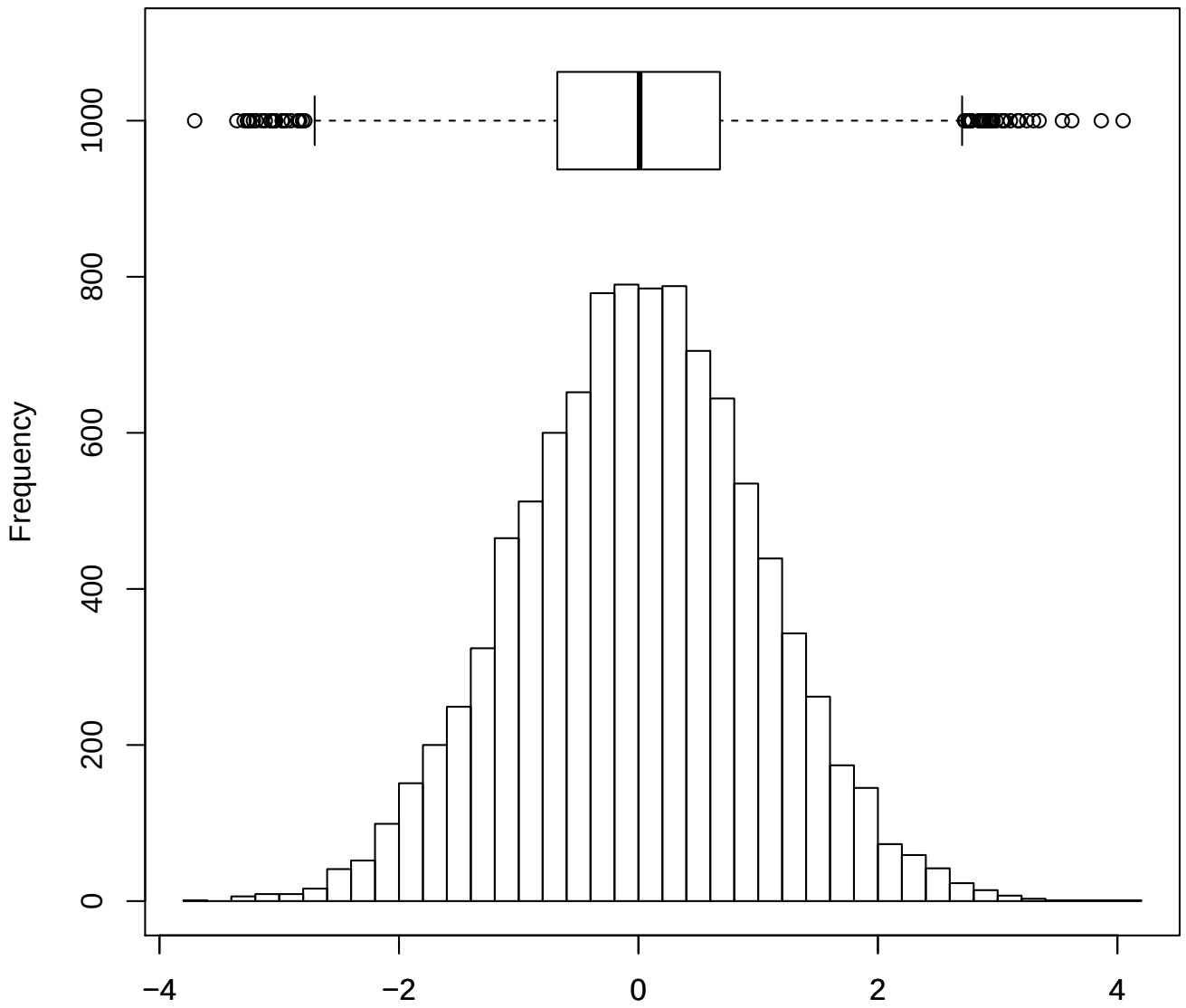

INGENIO·ET·LABORE

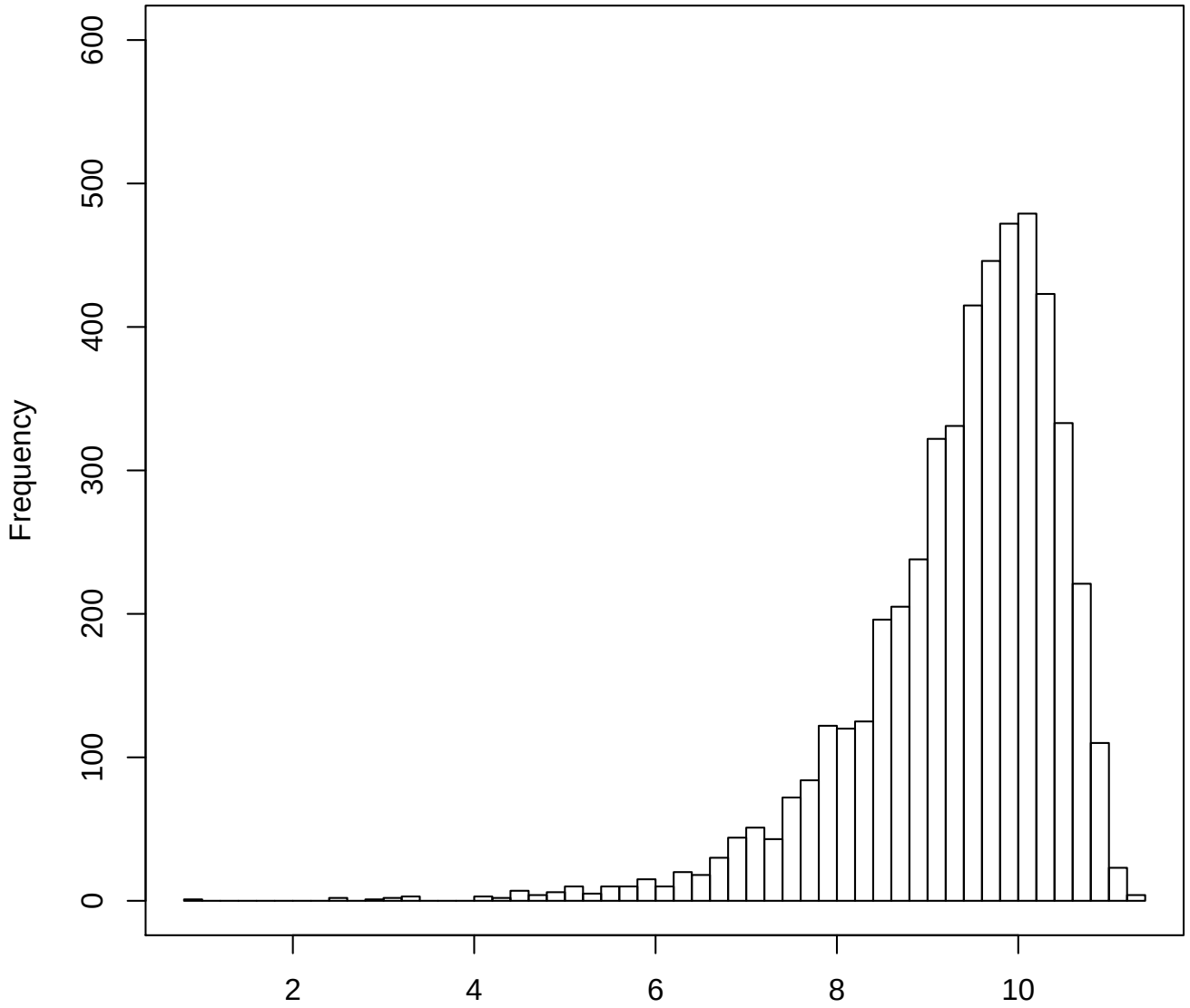

3.0
2.5
2.0
1.5
1.0
0.5
0.0

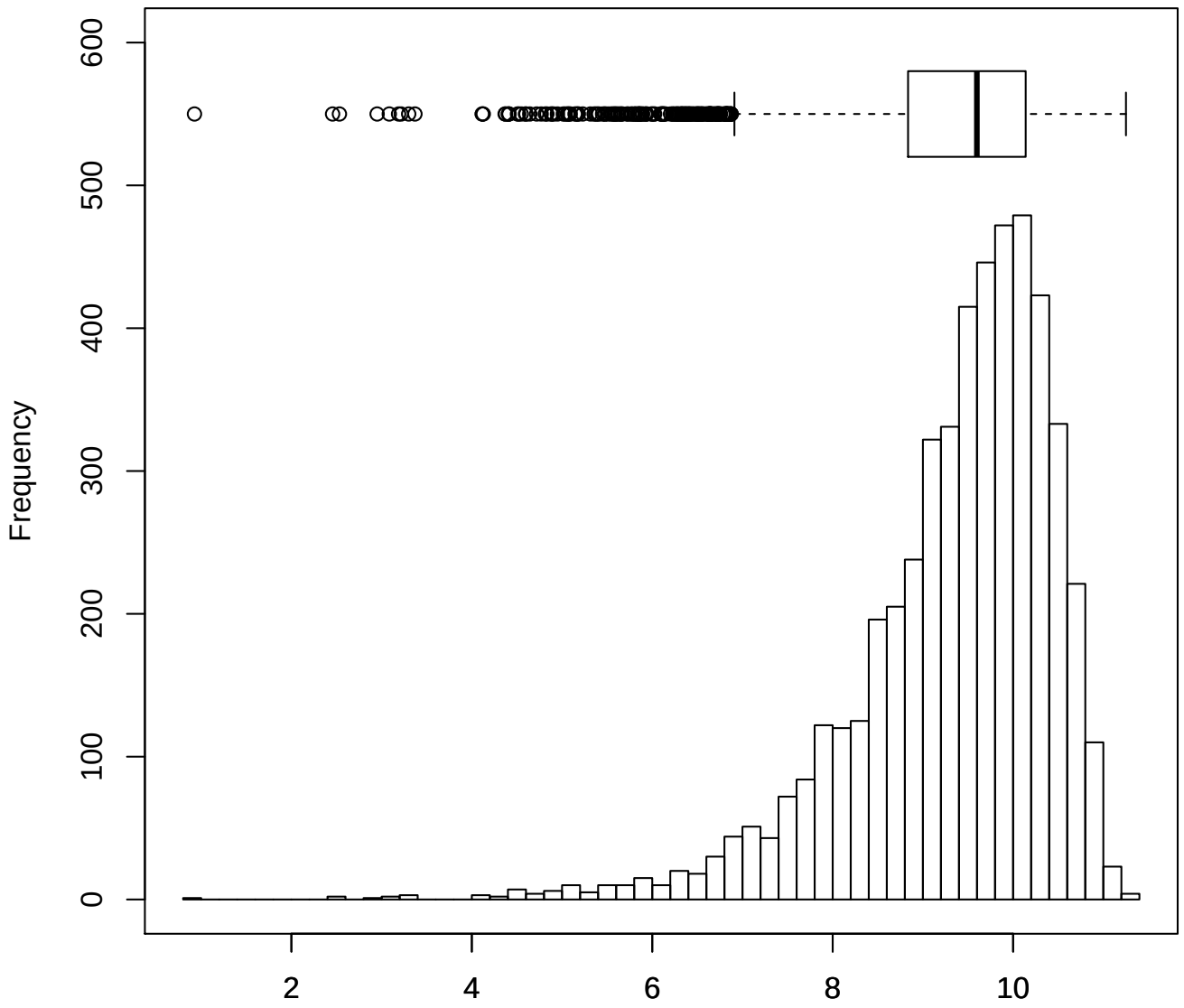
- Normal Quantile-Quantile (QQ) Plot

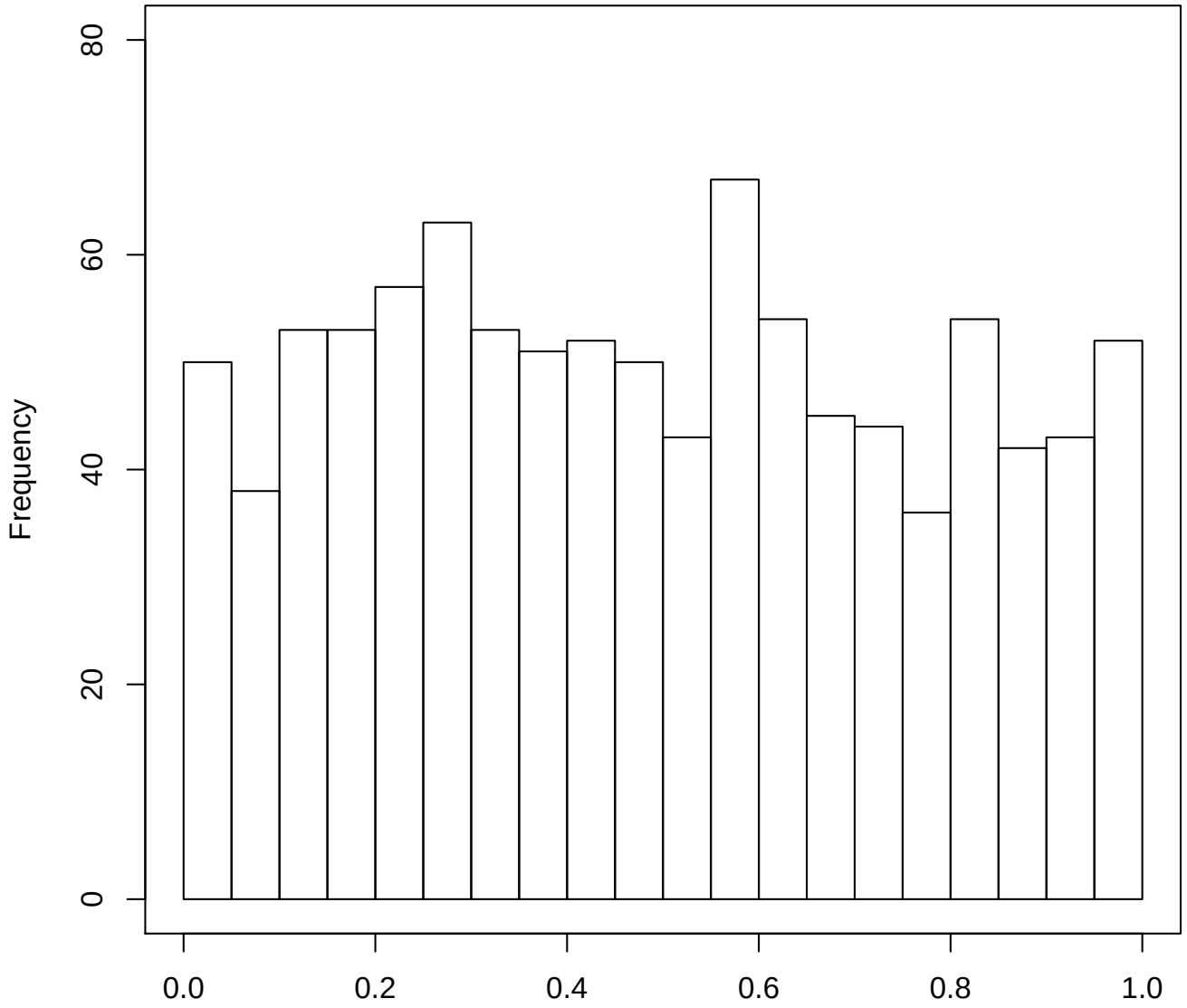


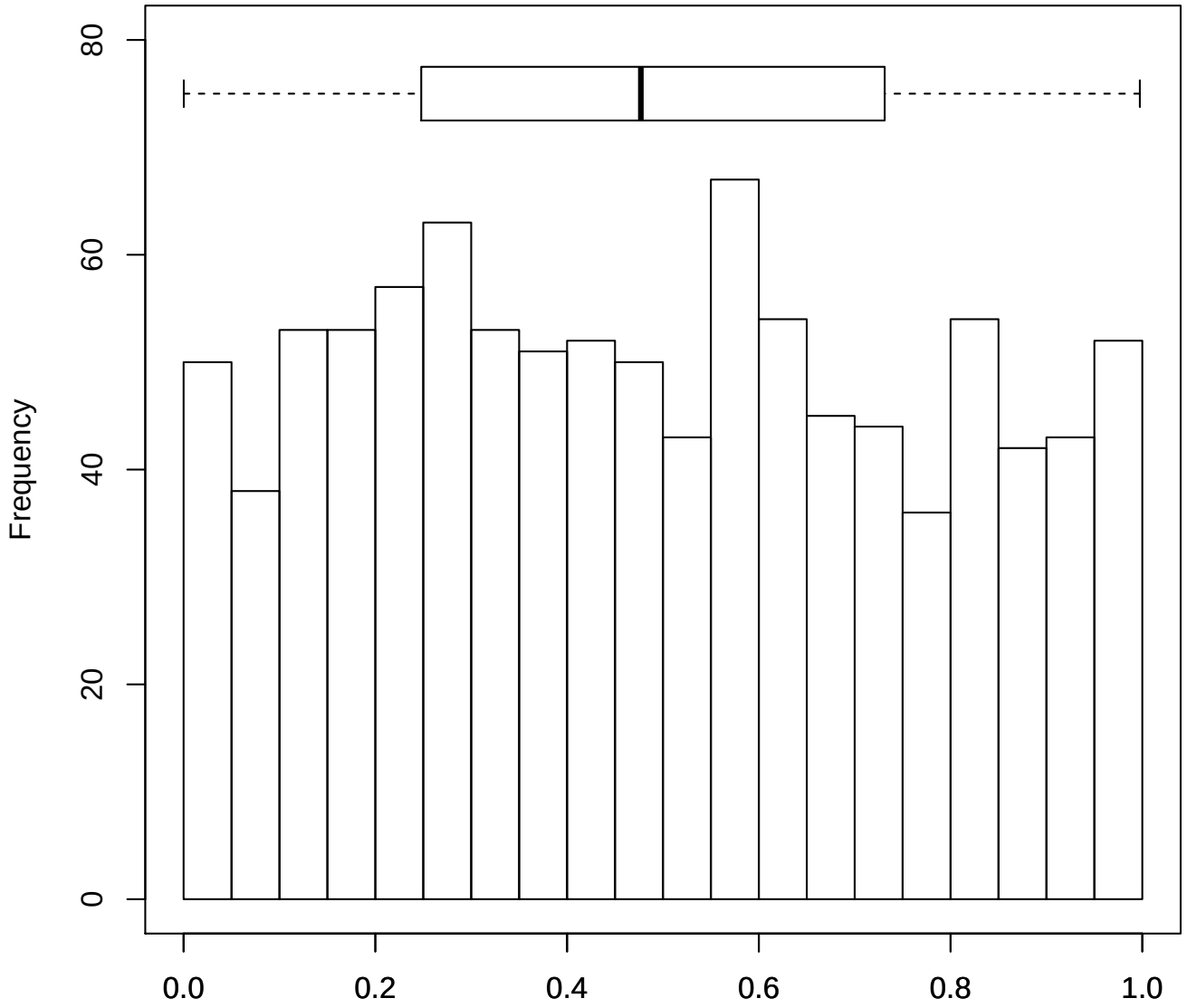

Histogram showing Normal Distribution
Frequency
0
200
400
600
800
1000
−4
−2
0
2
4


Histogram and boxplot showing Normal Distribution
Frequency
0
200
400
600
800
1000
−4
−2
0
2
4


Histogram showing left skewed distribution
Frequency
0
100
200
300
400
500
600
2
4
6
8
10


Histogram and boxplot showing left skewed istribution
Frequency
0
100
200
300
400
500
600
2
4
6
8
10


Histogram showing uniform distribution
Frequency
0
20
40
60
80
0.0
0.2
0.4
0.6
0.8
1.0


Histogram and boxplot showing uniform distribution
Frequency
0
20
40
60
80
0.0
0.2
0.4
0.6
0.8
1.0

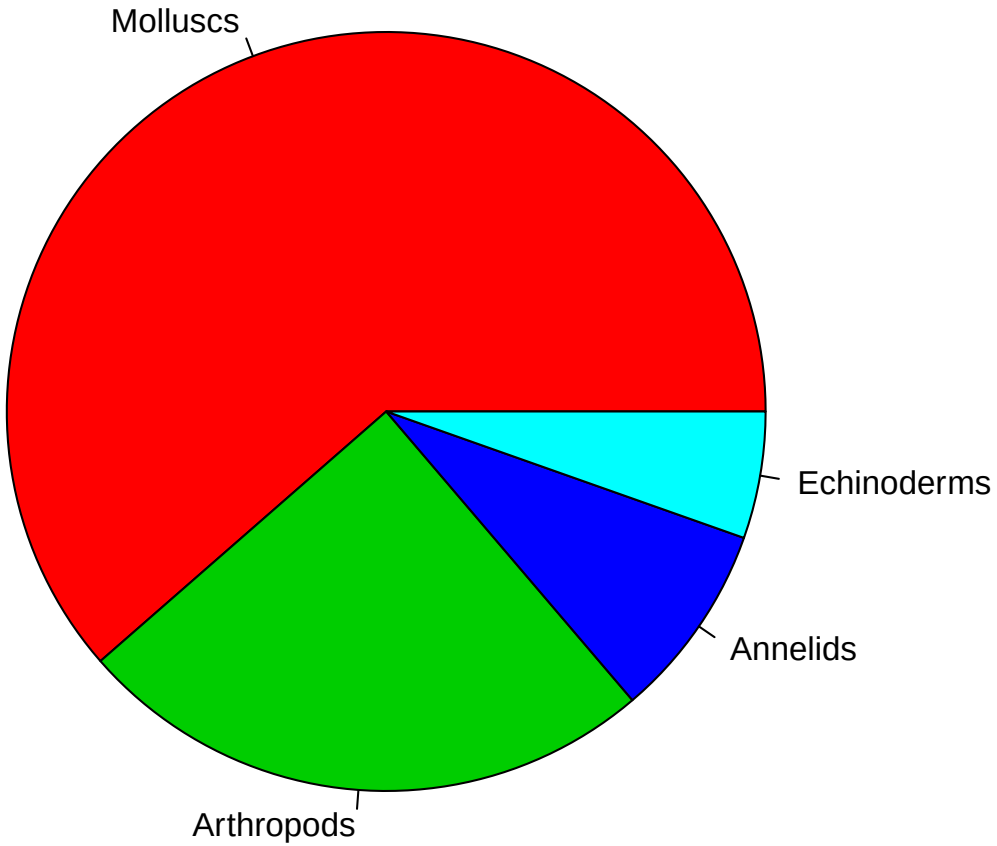
One Qualitative Variable

- Frequency table

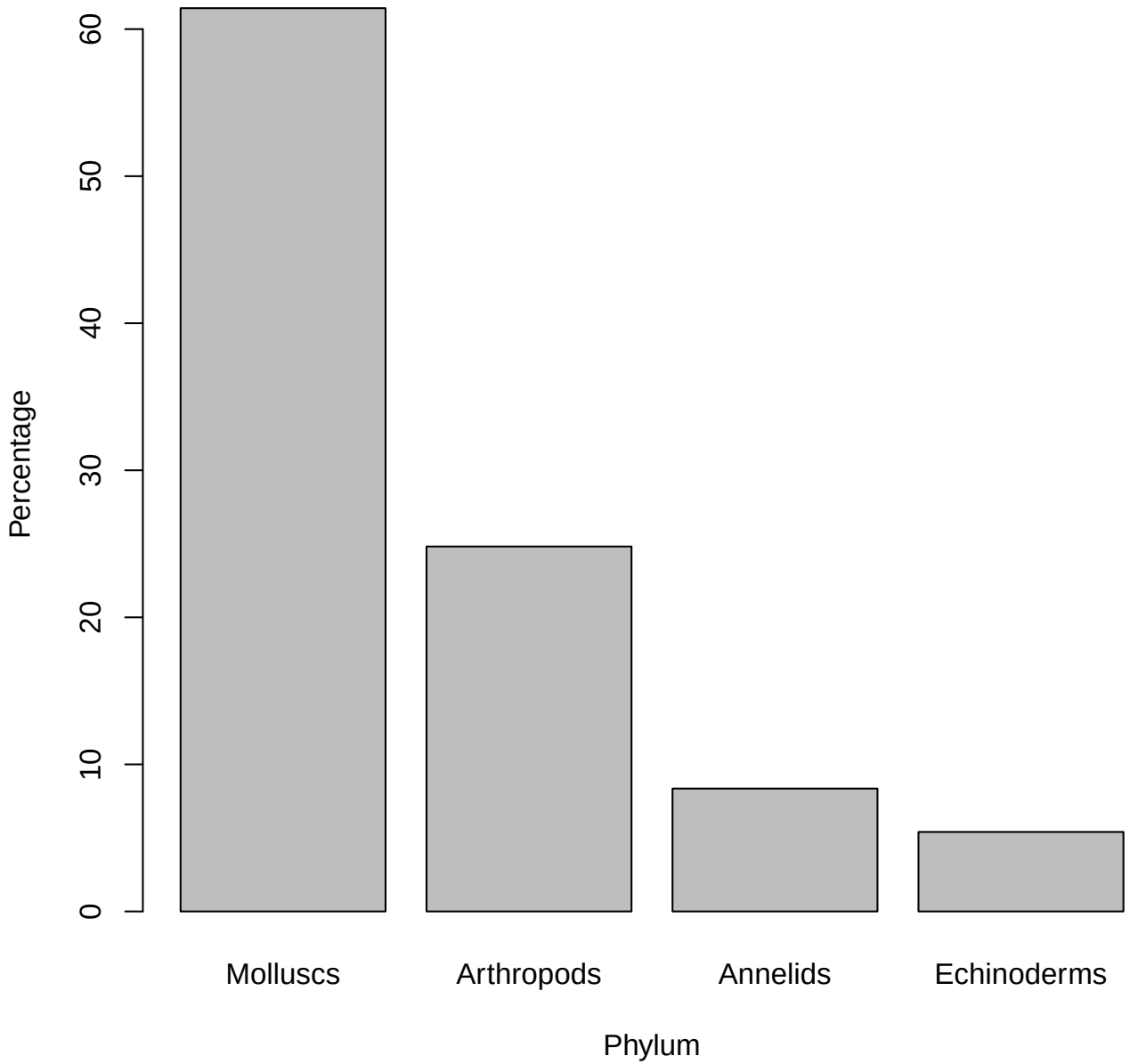
Table: One way frequency table

Phylum	Frequency	Percentage
Molluscs	250	61.4%
Annelids	34	8.4%
Arthropods	101	24.8%
Echinoderms	22	5.4%
Total	407	100%

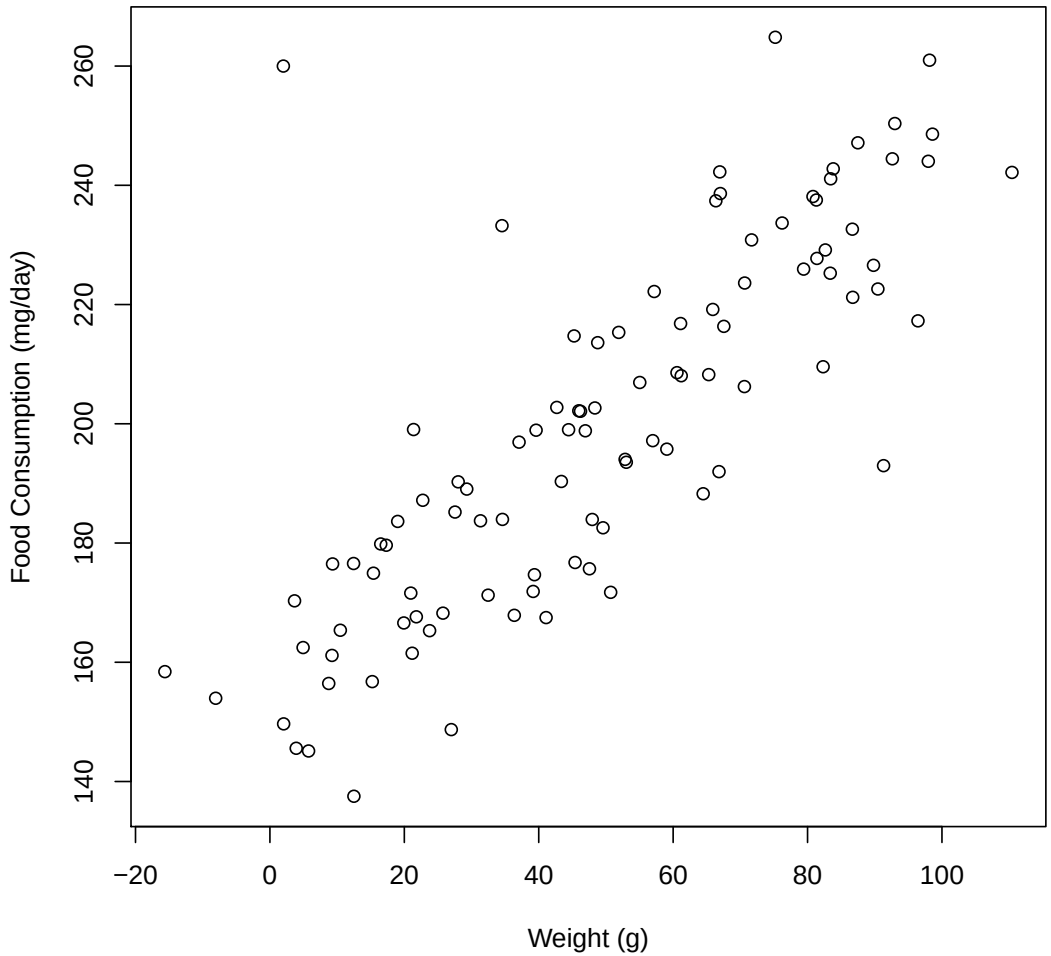
- Pie chart
- Barplot


Molluscs
Echinoderms
Annelids
Arthropods

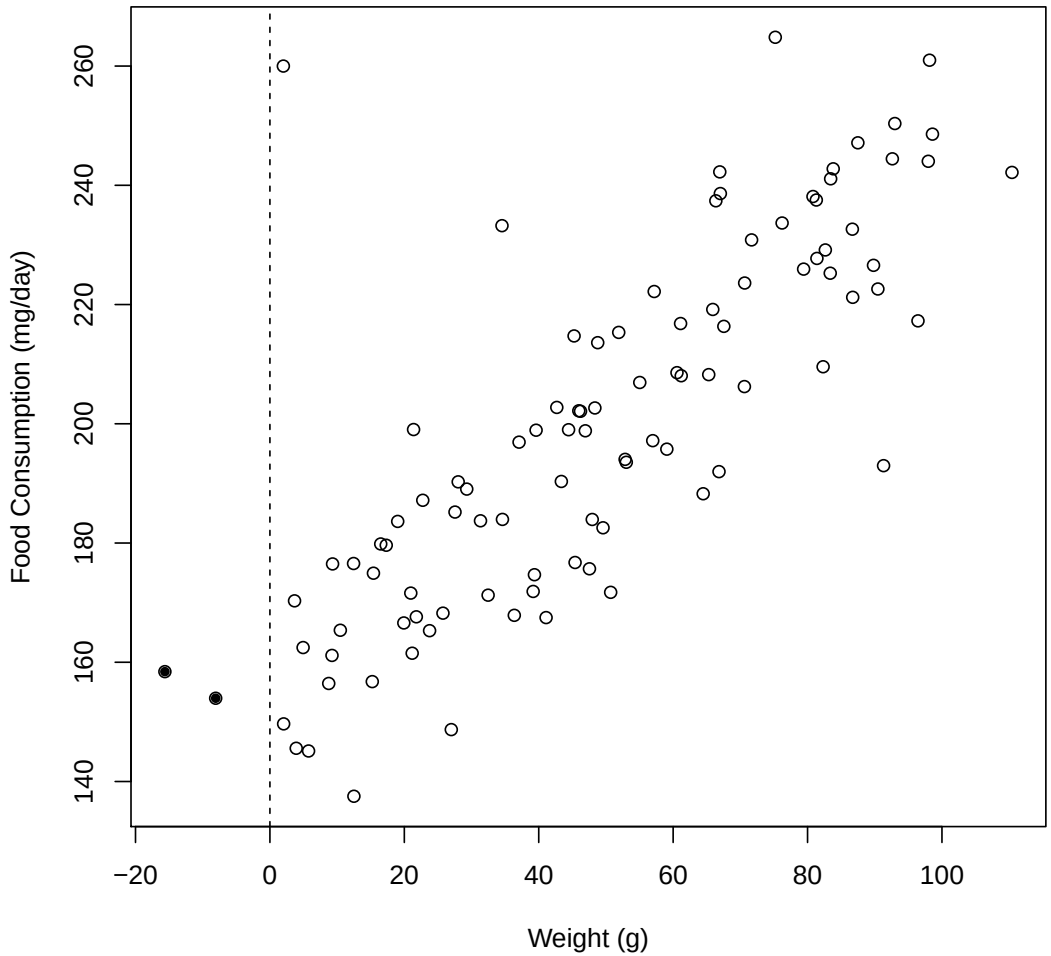

INGENIO·ET·LABORE


Percentage
0
10
20
30
40
50
60
Molluscs
Arthropods
Annelids
Echinoderms
Phylum

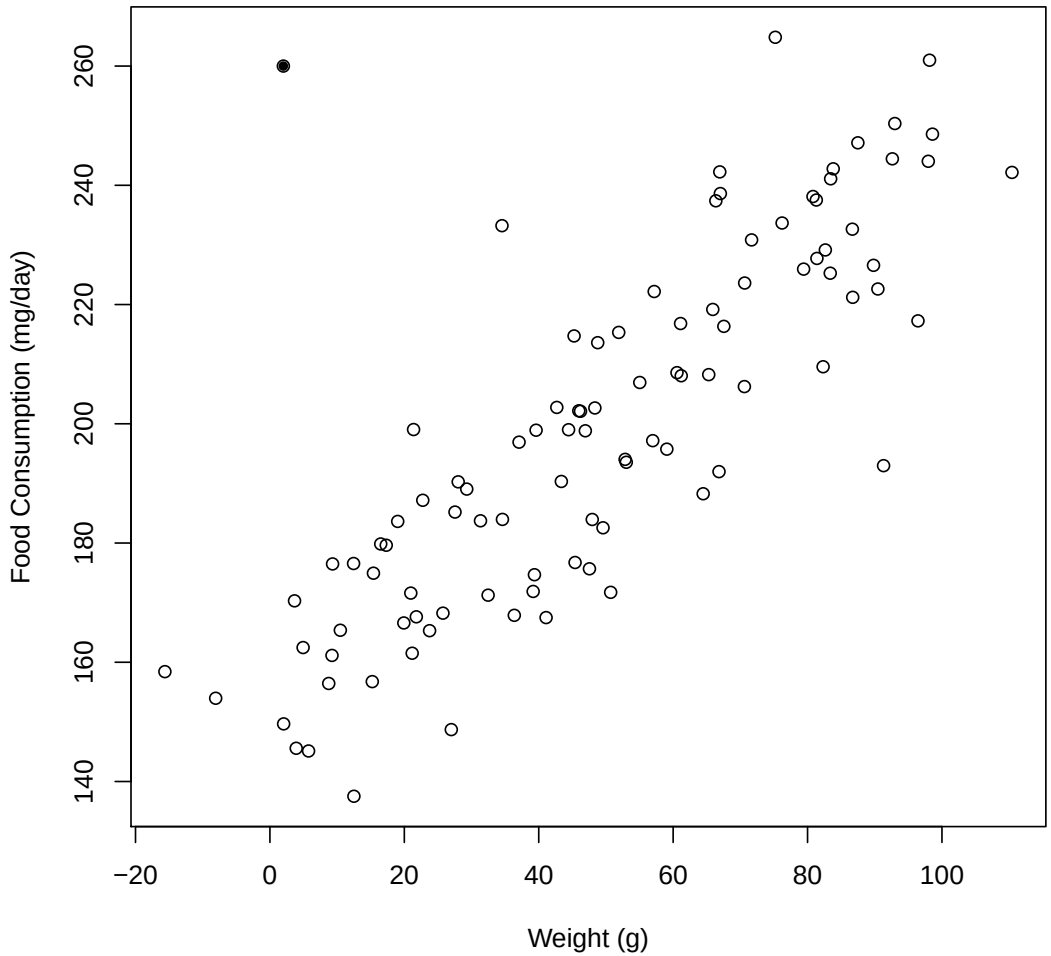
- Scatter Plot



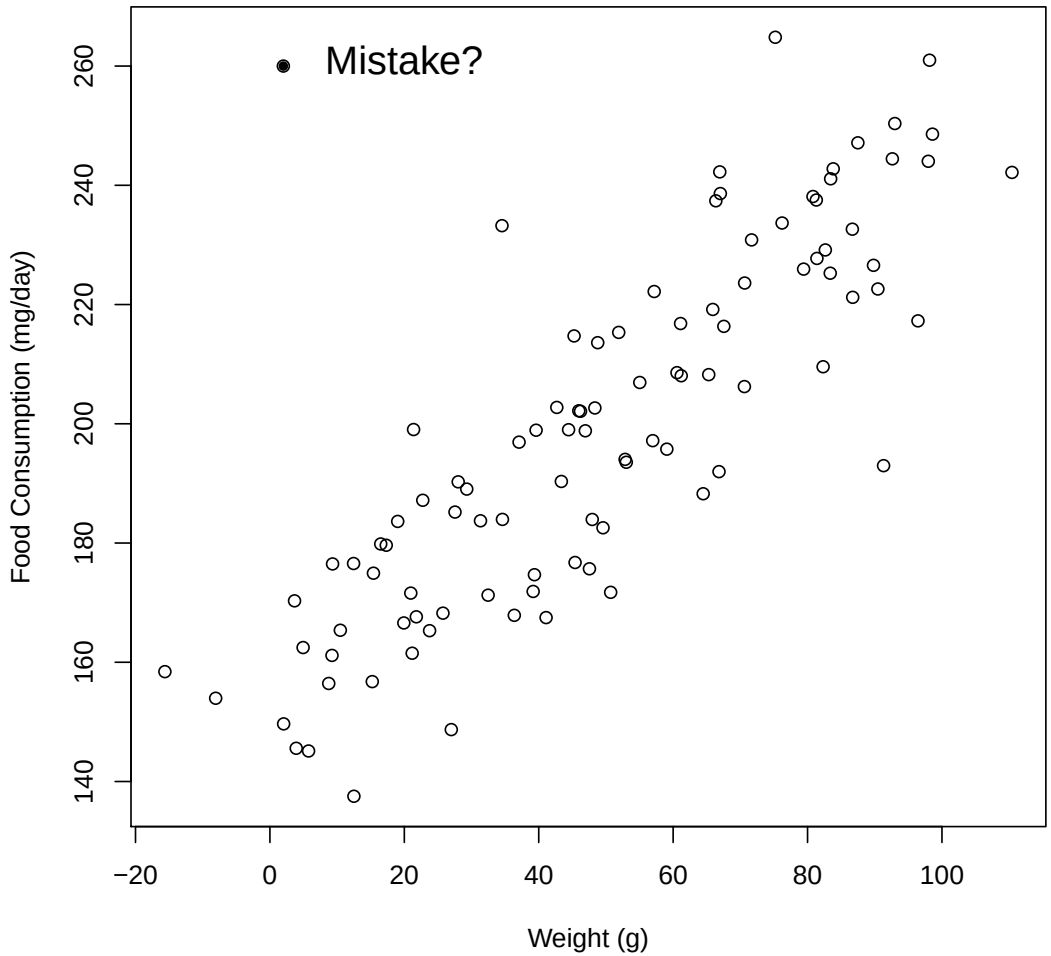
- Scatter Plot



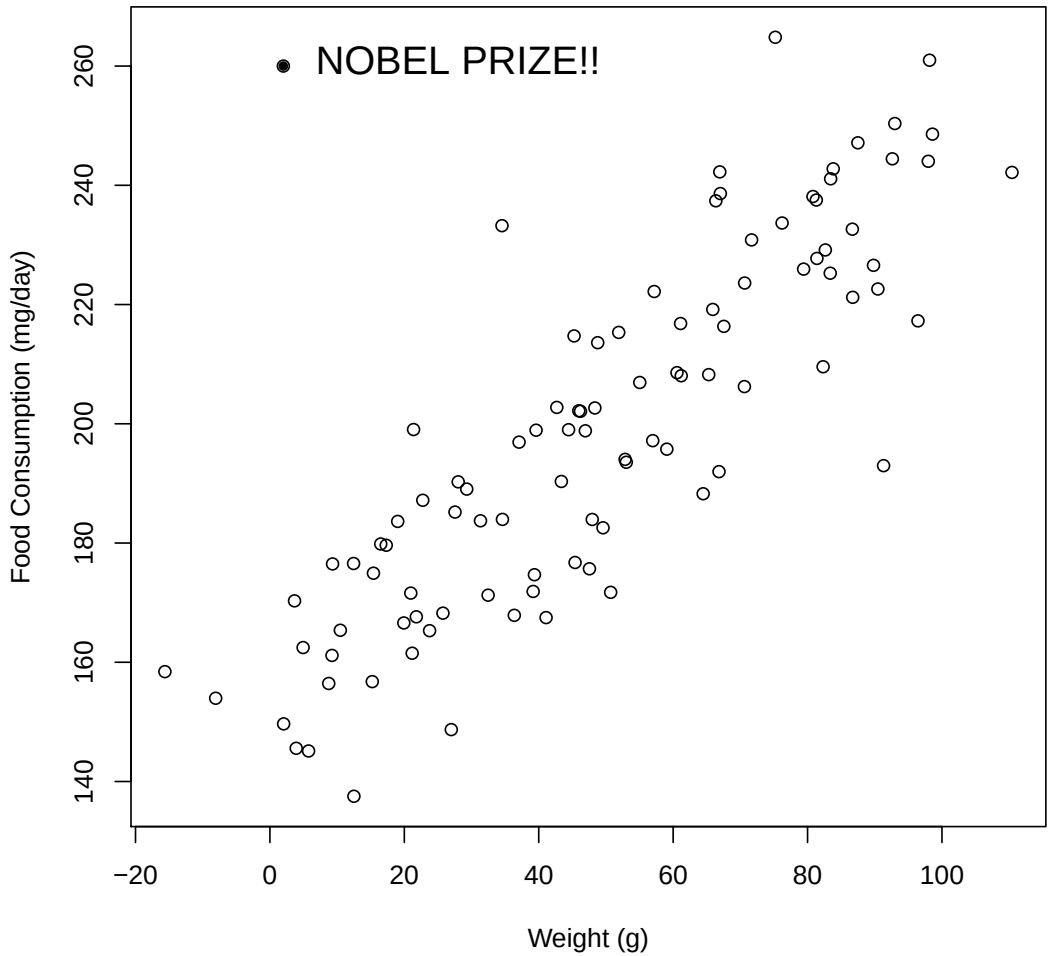
- Scatter Plot



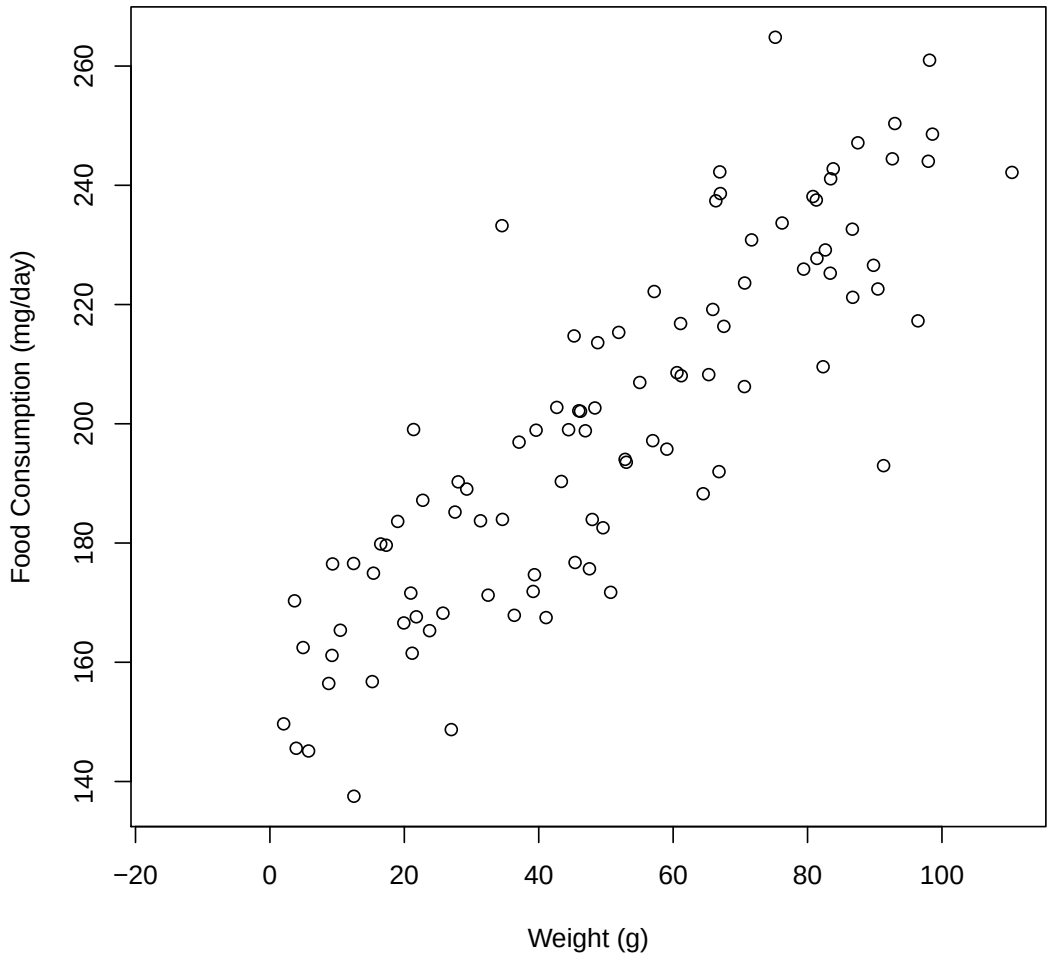
- Scatter Plot



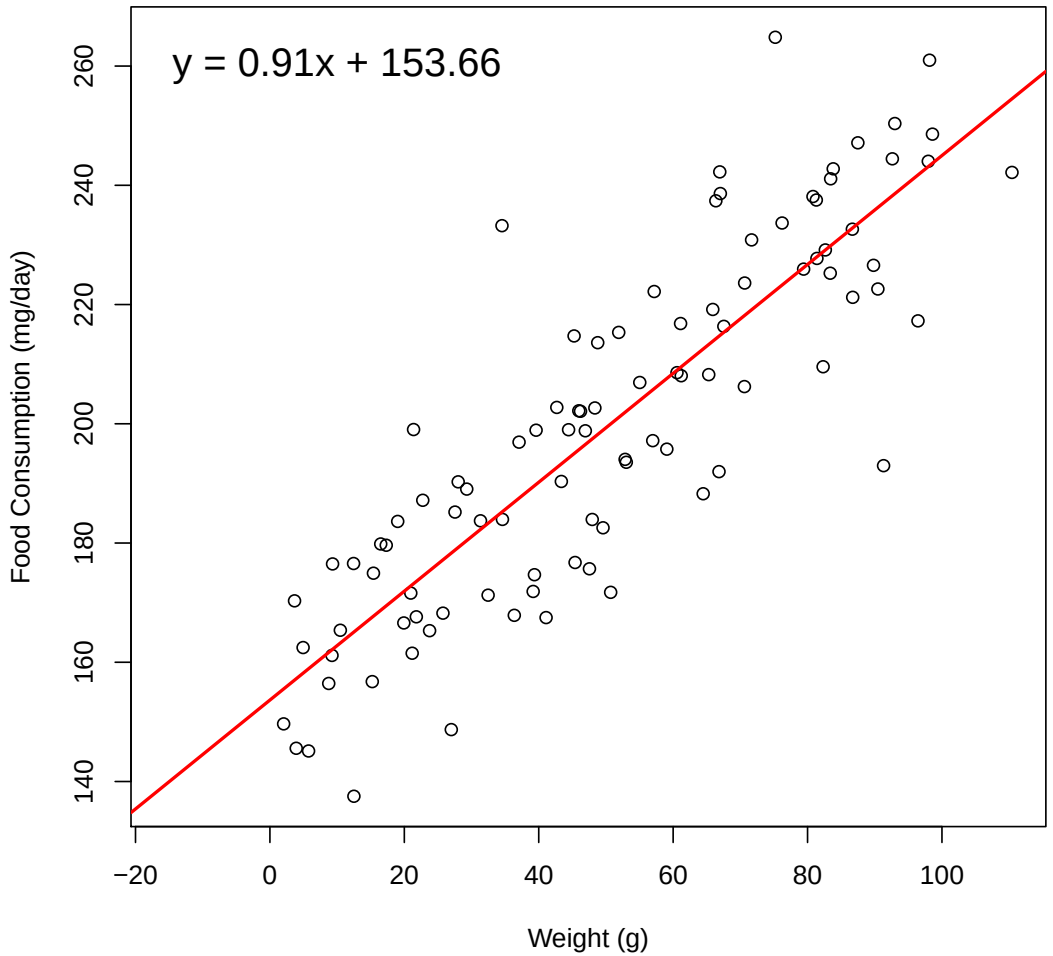
- Scatter Plot



- Scatter Plot



- Scatter Plot

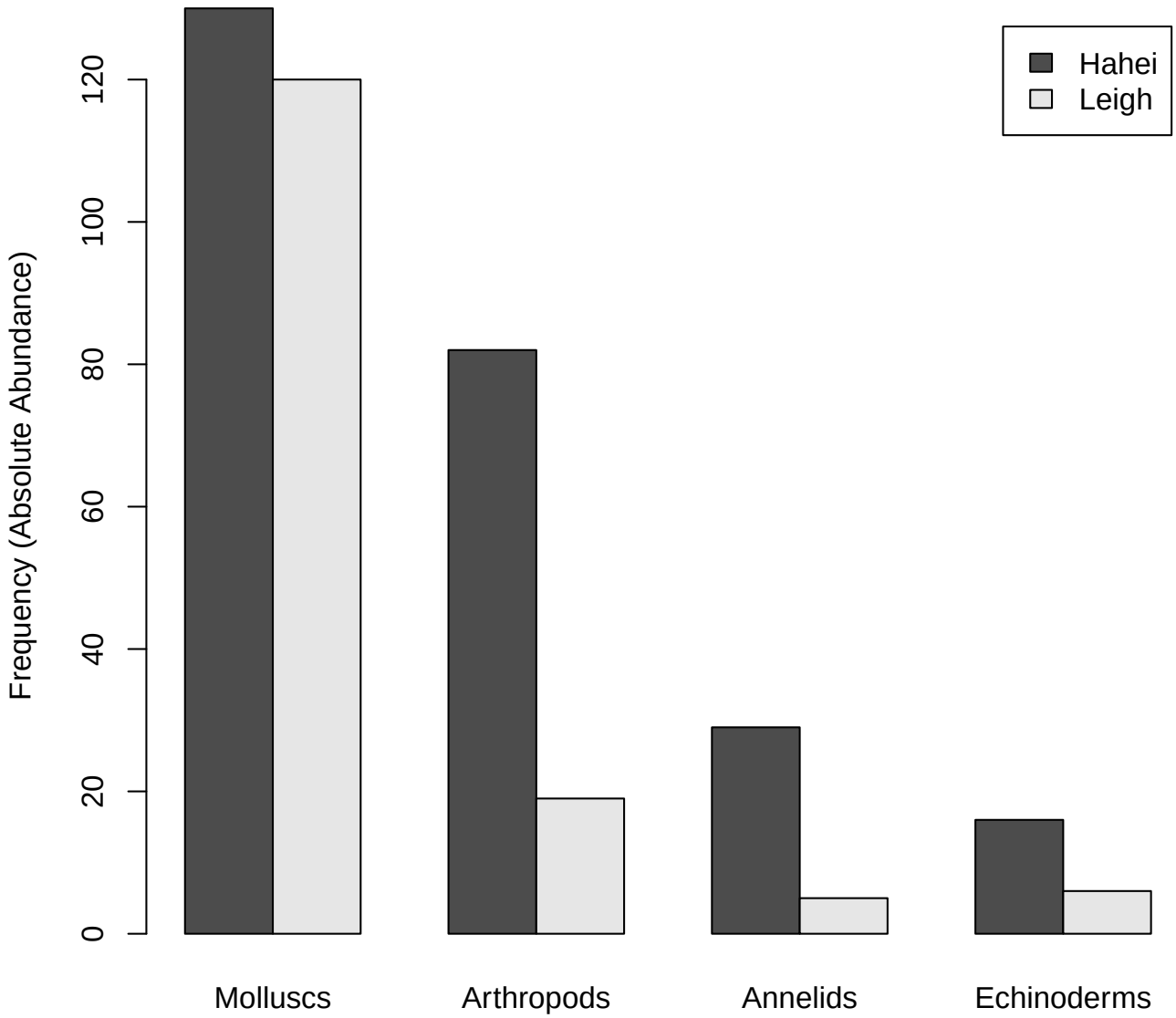


- Two way frequency table

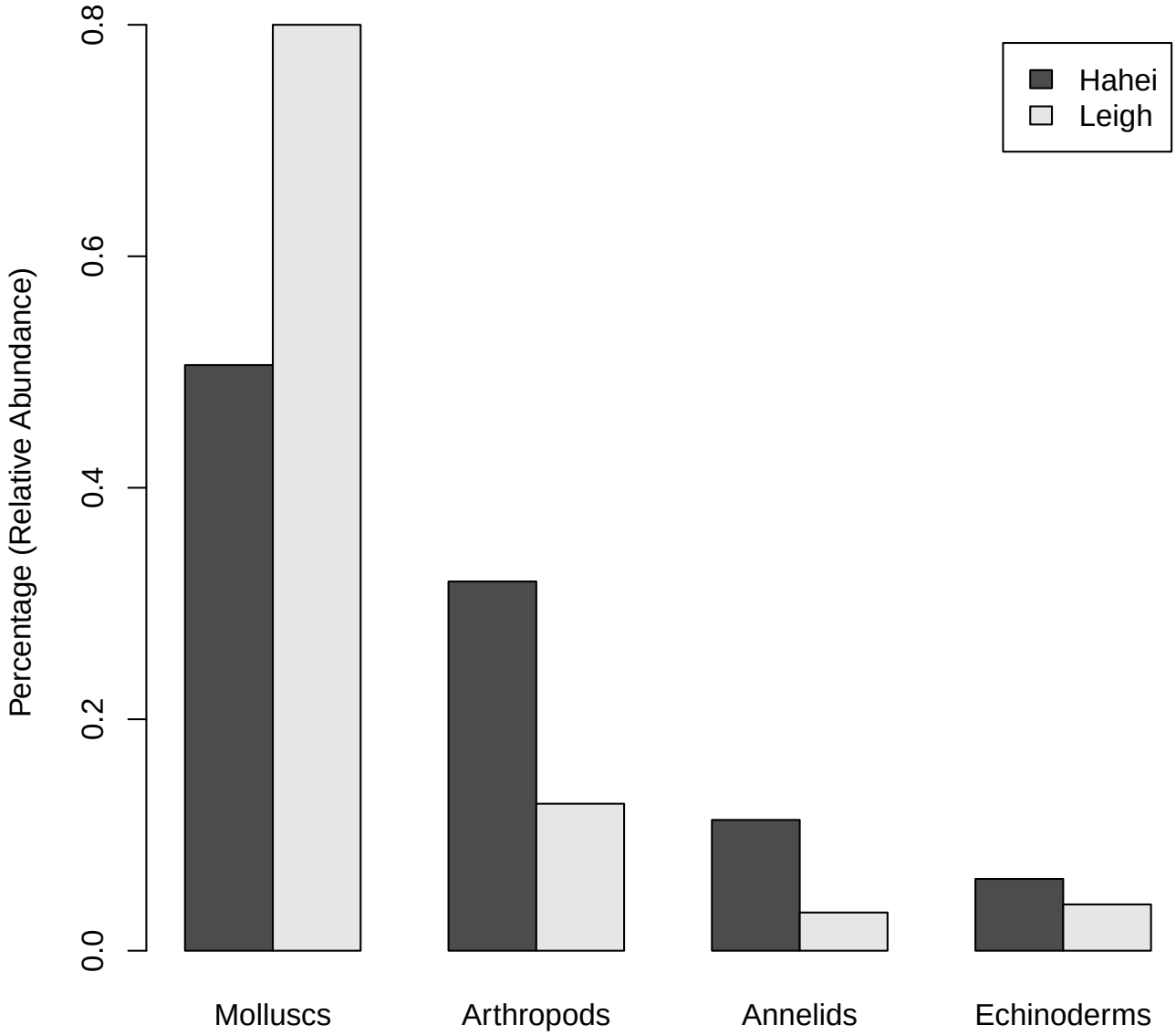
Table: Two way frequency table

Phylum	Hahei	Leigh	Total
Molluscs	130	120	250
Annelids	29	5	34
Arthropods	82	19	101
Echinoderms	16	6	22
Total	257	150	407

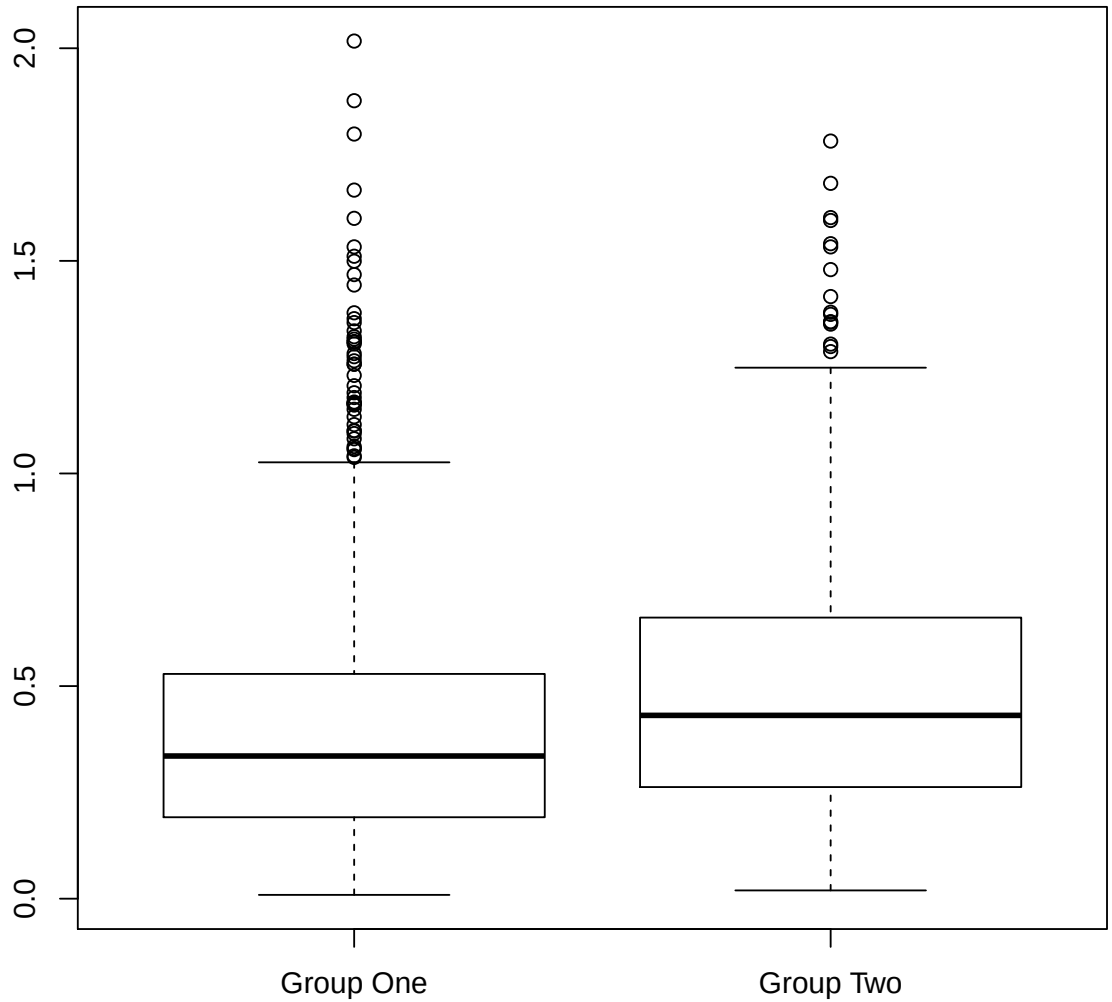
- Side by side barplot: frequencies or percentages?


Hahei
Leigh
Frequency (Absolute Abundance)
0
20
40
60
80
100
120
Molluscs
Arthropods
Annelids
Echinoderms

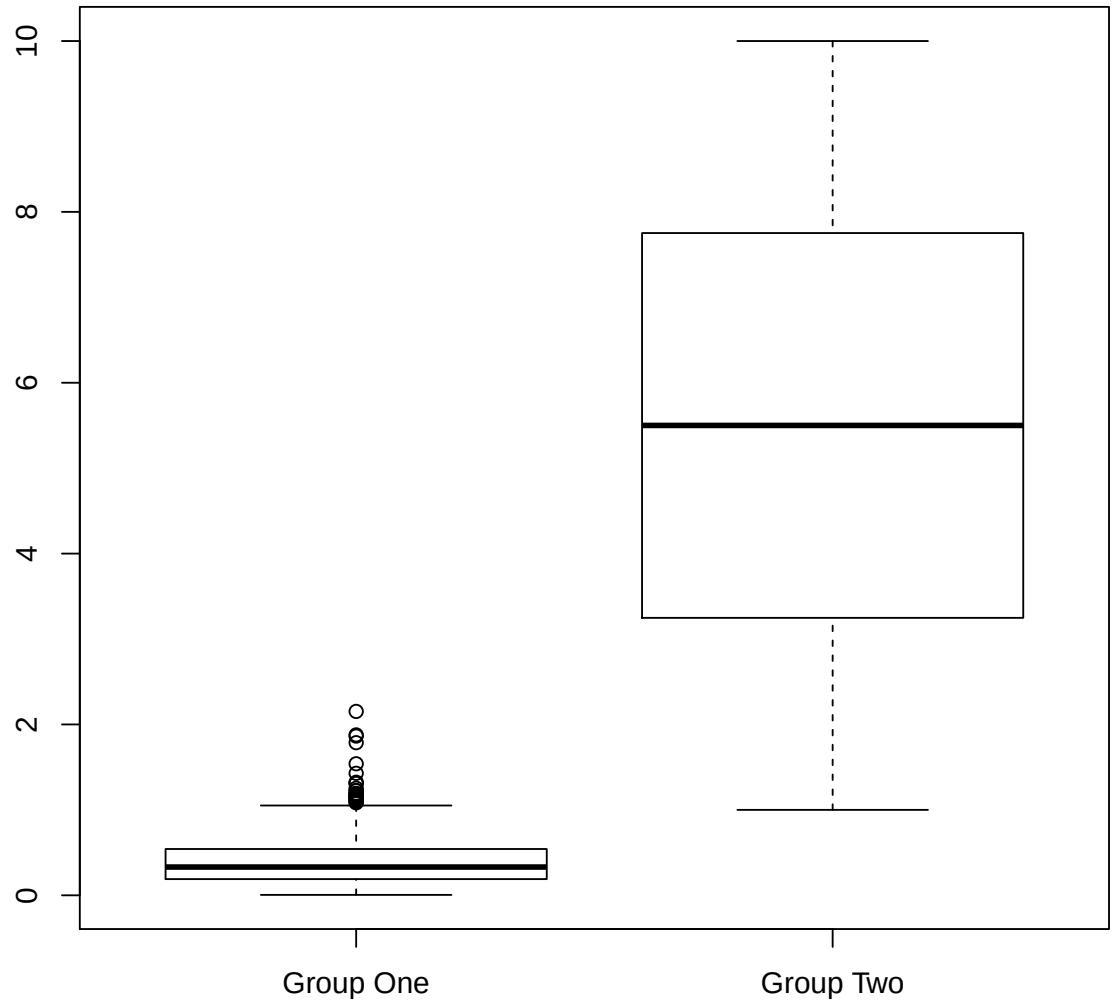

INGENIO·ET·LABORE


Percentage (Relative Abundance)
0.0
0.2
0.4
0.6
0.8
Hahei
Leigh
Molluscs
Arthropods
Annelids
Echinoderms

- Side-by-side boxplot



- Side-by-side boxplot



- QQ Plot

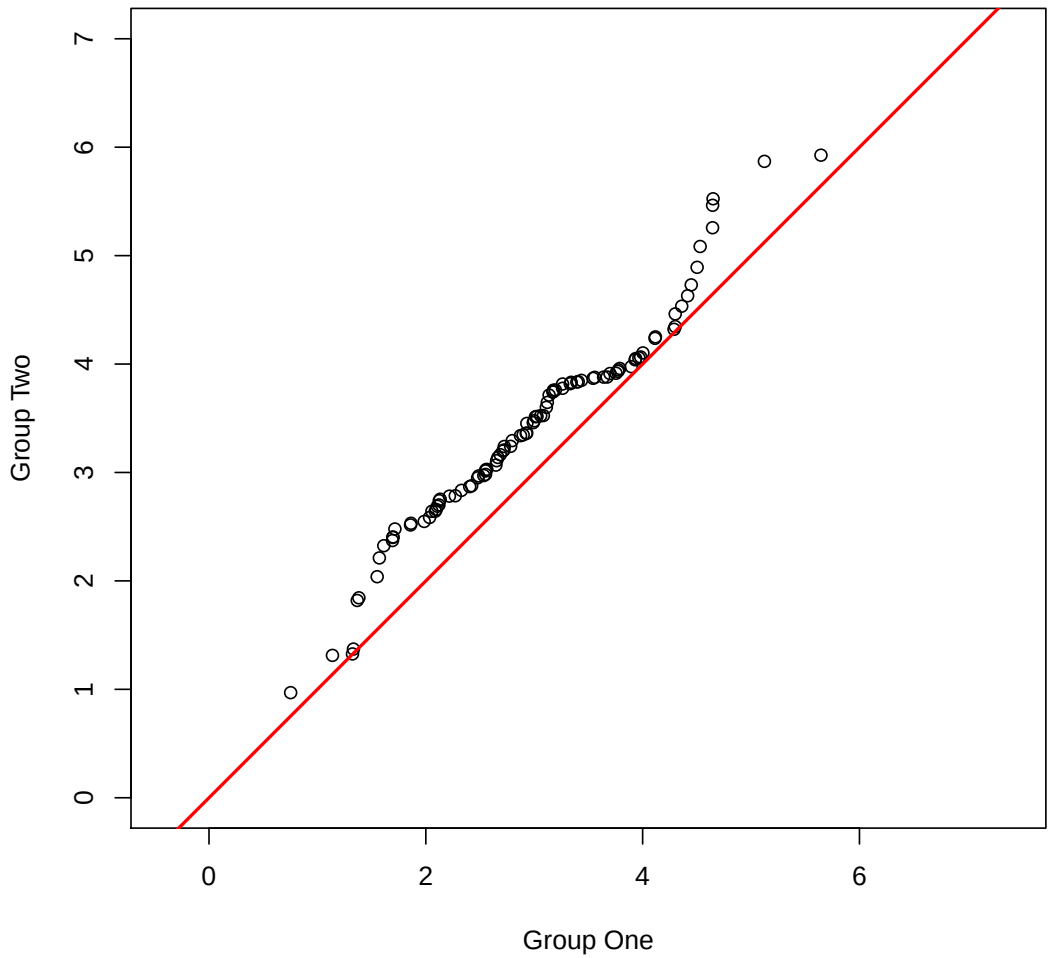
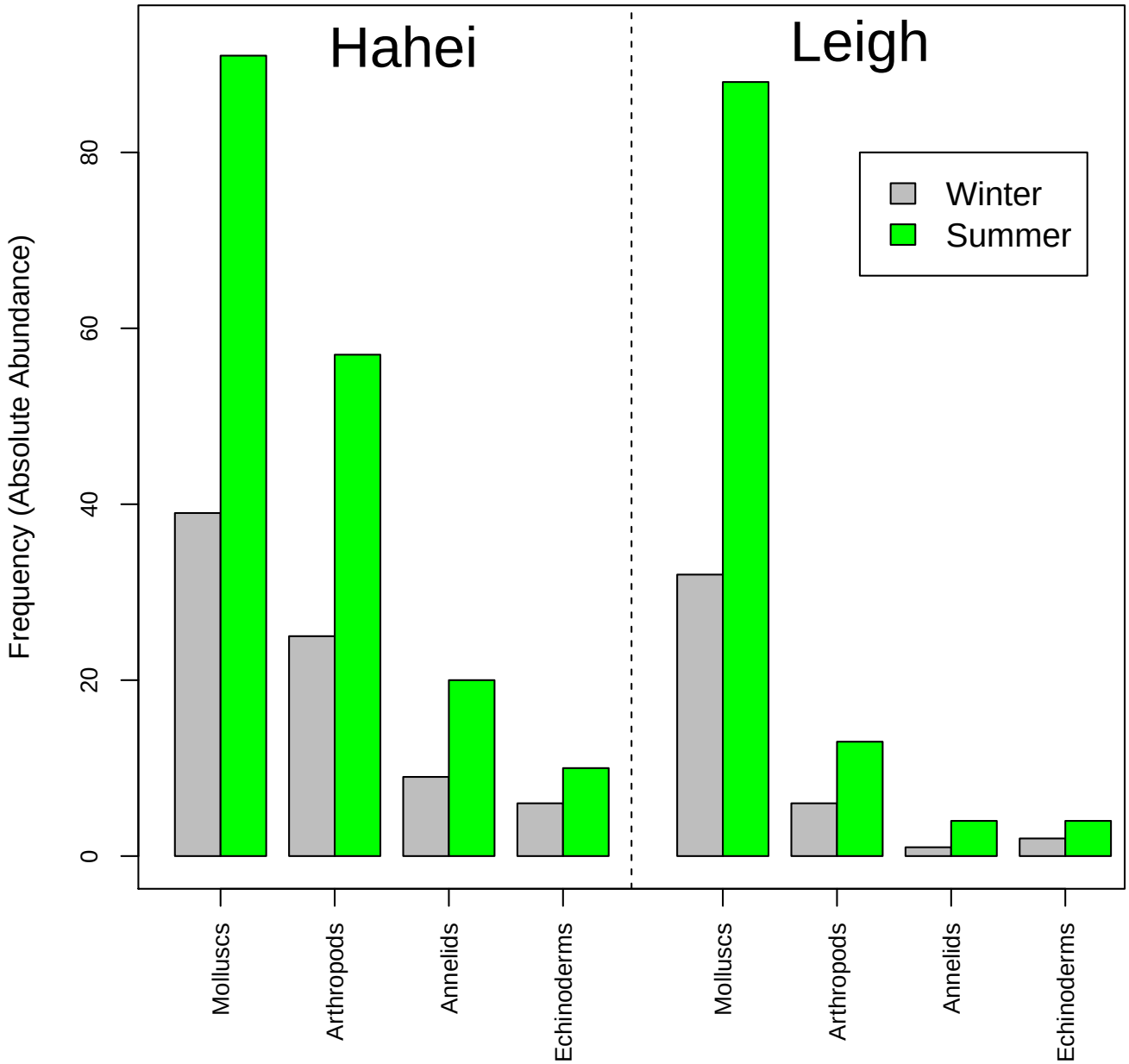
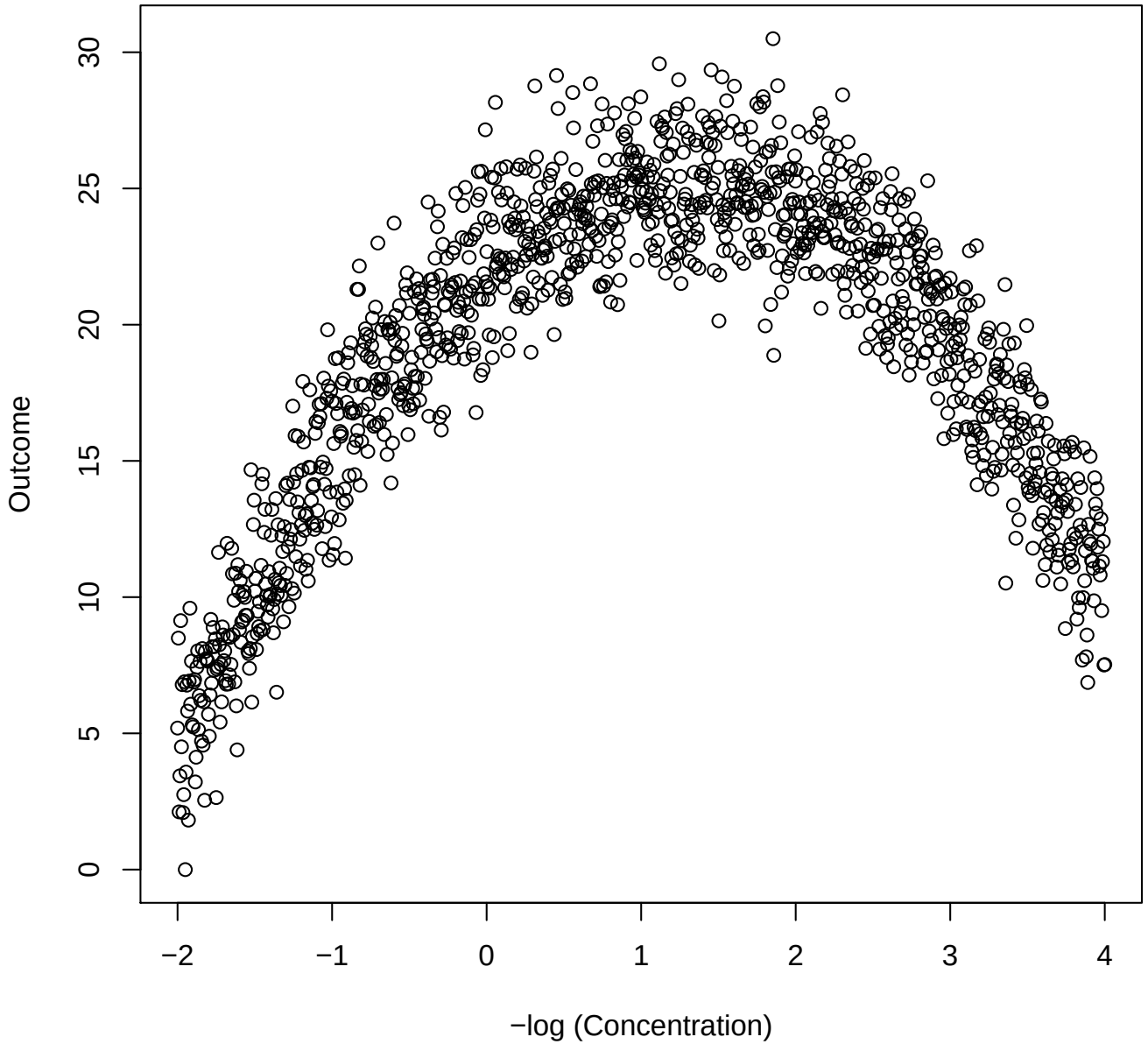


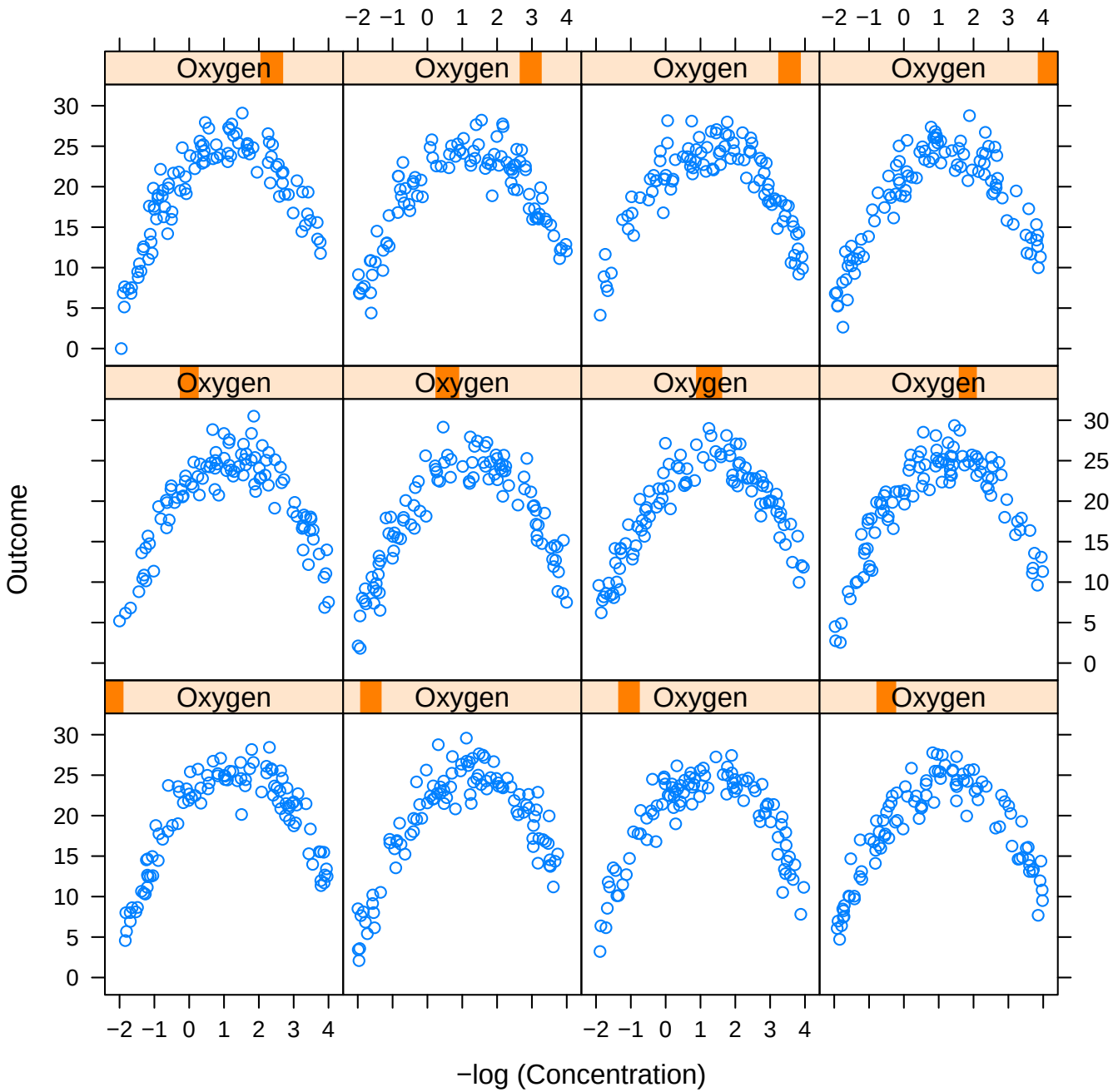
Table: Three way frequency table

	Location			
	Hahei		Leigh	
Phylum	Winter	Summer	Winter	Summer
Molluscs	39	91	32	88
Annelids	9	20	1	4
Arthropods	25	57	6	13
Echinoderms	6	10	2	4


Hahei
Leigh
Winter
Summer
Frequency (Absolute Abundance)
0
20
40
60
80
Molluscs
Arthropods
Annelids
Echinoderms
Molluscs
Arthropods
Annelids
Echinoderms

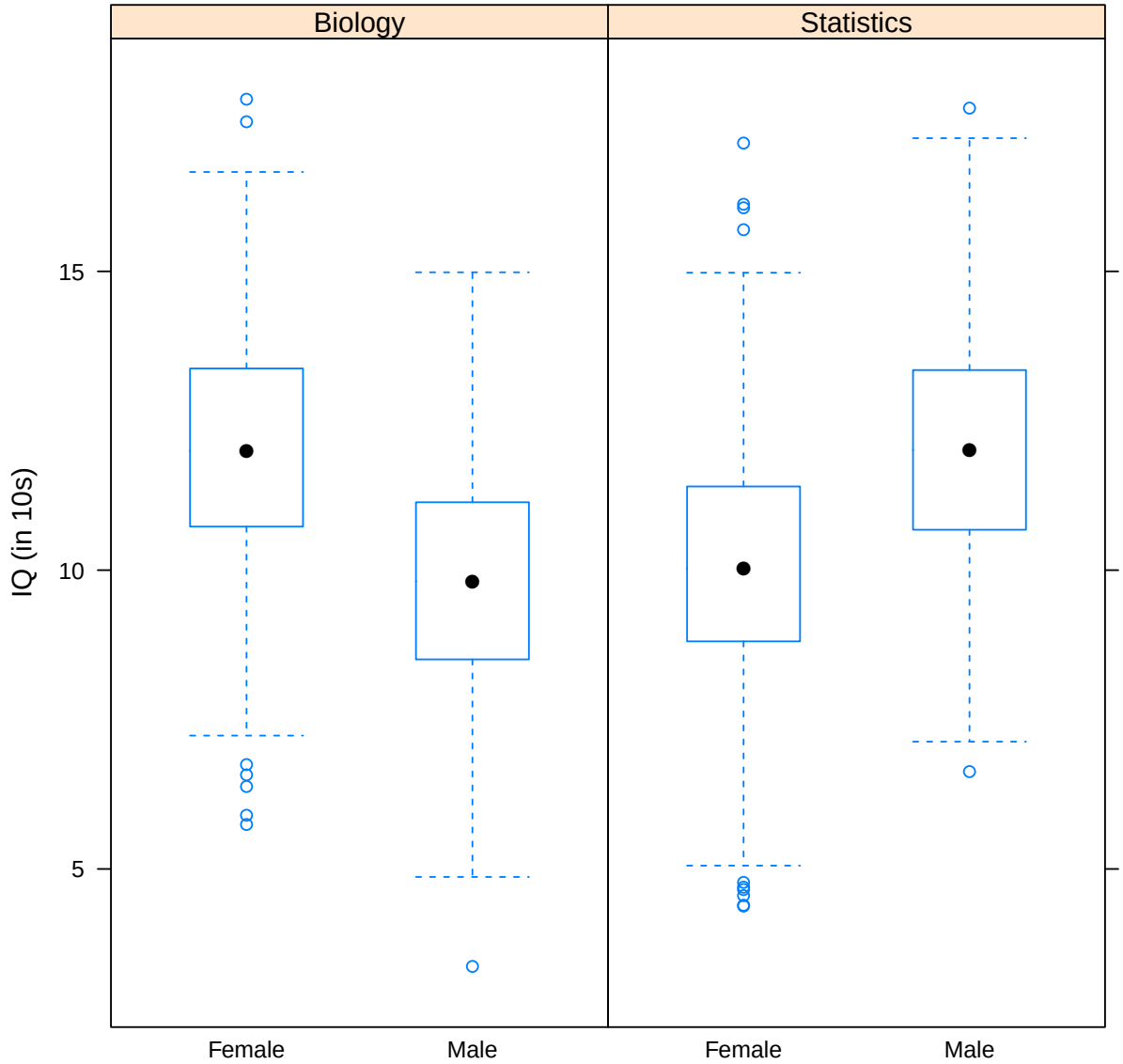
- Trellis Graphs: Display a variable or the relationship between variables, conditioned on one or more other variables.
  - Three continuous variables *-log(Concentration)*, *Outcome* and *Oxygen level*.
  - What is the relationship between *-log(Concentration)* and *Outcome*, conditioned on *Oxygen level*? In other words, how does the relationship between *-log(Concentration)* and *Outcome* change over different *Oxygen level*?
    1. Plot *-log(Concentration)* against *Outcome* first.
    2. Break *Oxygen level* into ordinal groups. e.g. Divide *Oxygen level* into 12 equally spaced interval.


Outcome
0
5
10
15
20
25
30
−2
−1
0
1
2
3
4
−log (Concentration)


Oxygen
Outcome
−log (Concentration)

Three Variables

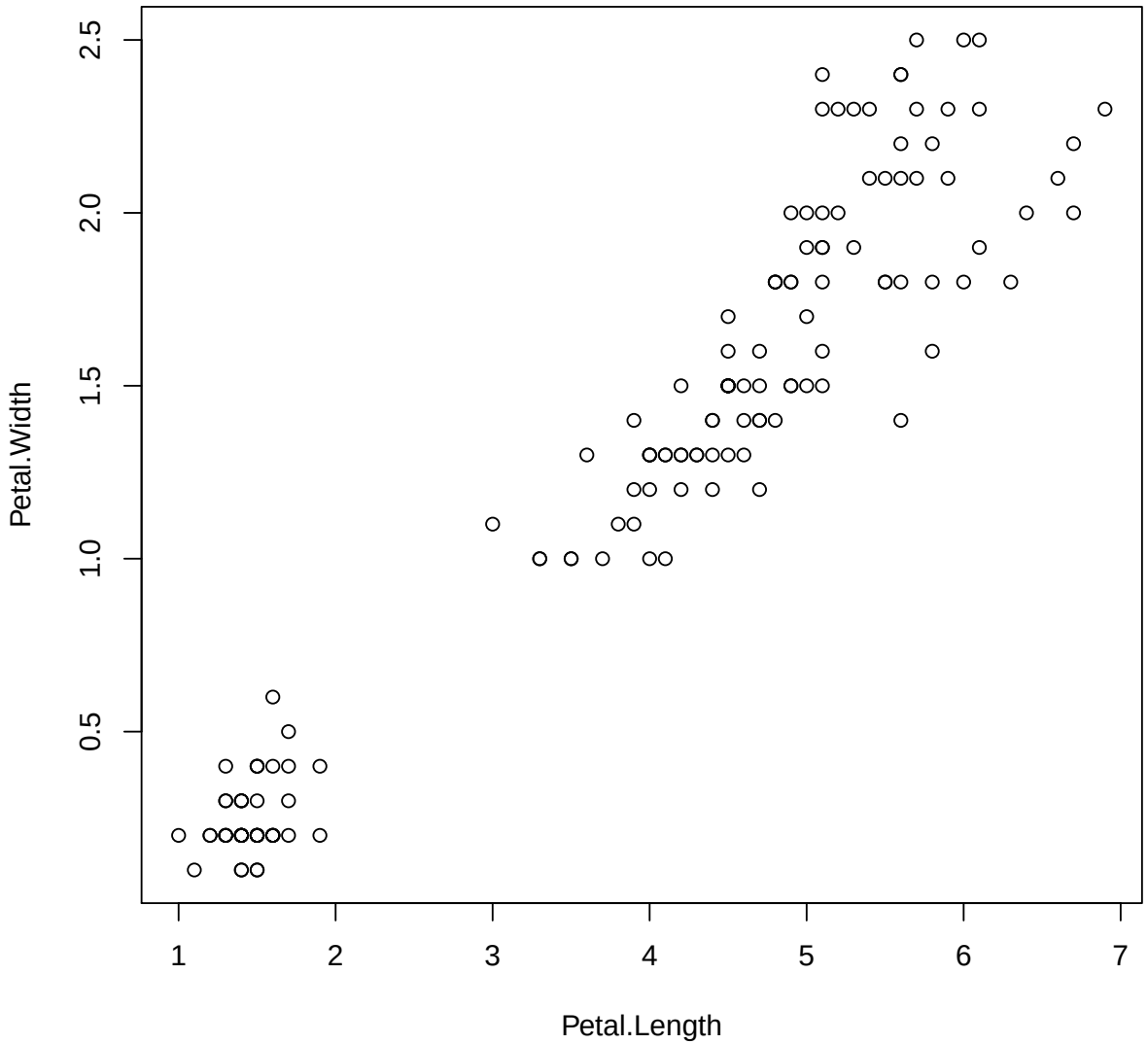
Two Qualitative Variables and One Quantitative Variable



- *Iris* flower data set
- Four continuous measure on Petal Length, Petal Width, Sepal Length and Sepal Width.
- One nominal variable, Species.
- Suppose we are interested in the relationship between Petal Length and Petal Width, and how such relationship changes for different species.
  1. Plot Petal Length against Petal Width.
  2. Use different plotting character/color to represent different species.

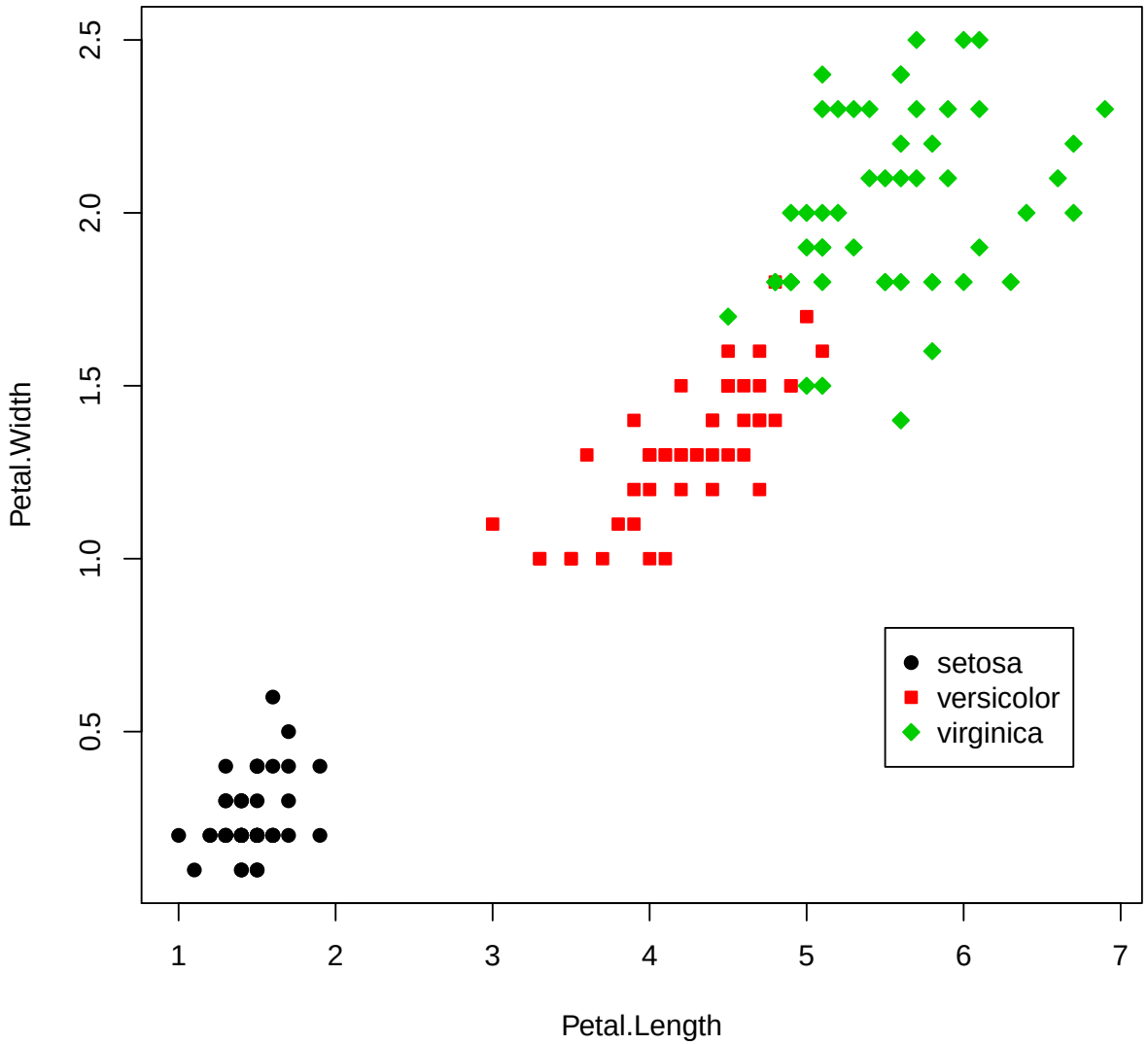
Three Variables

Two Quantitative Variables and One Qualitative Variable



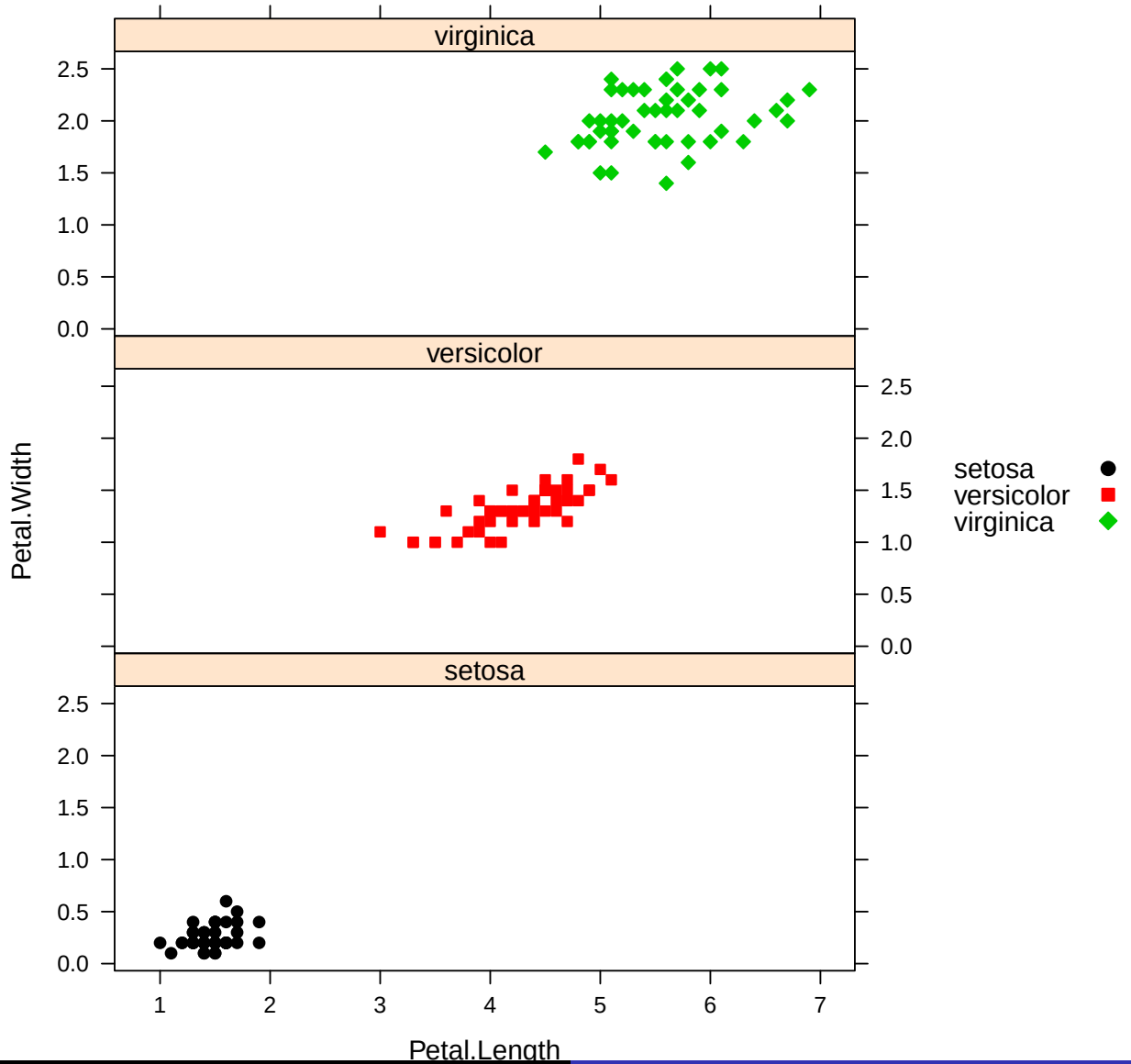
Three Variables

Two Quantitative Variables and One Qualitative Variable



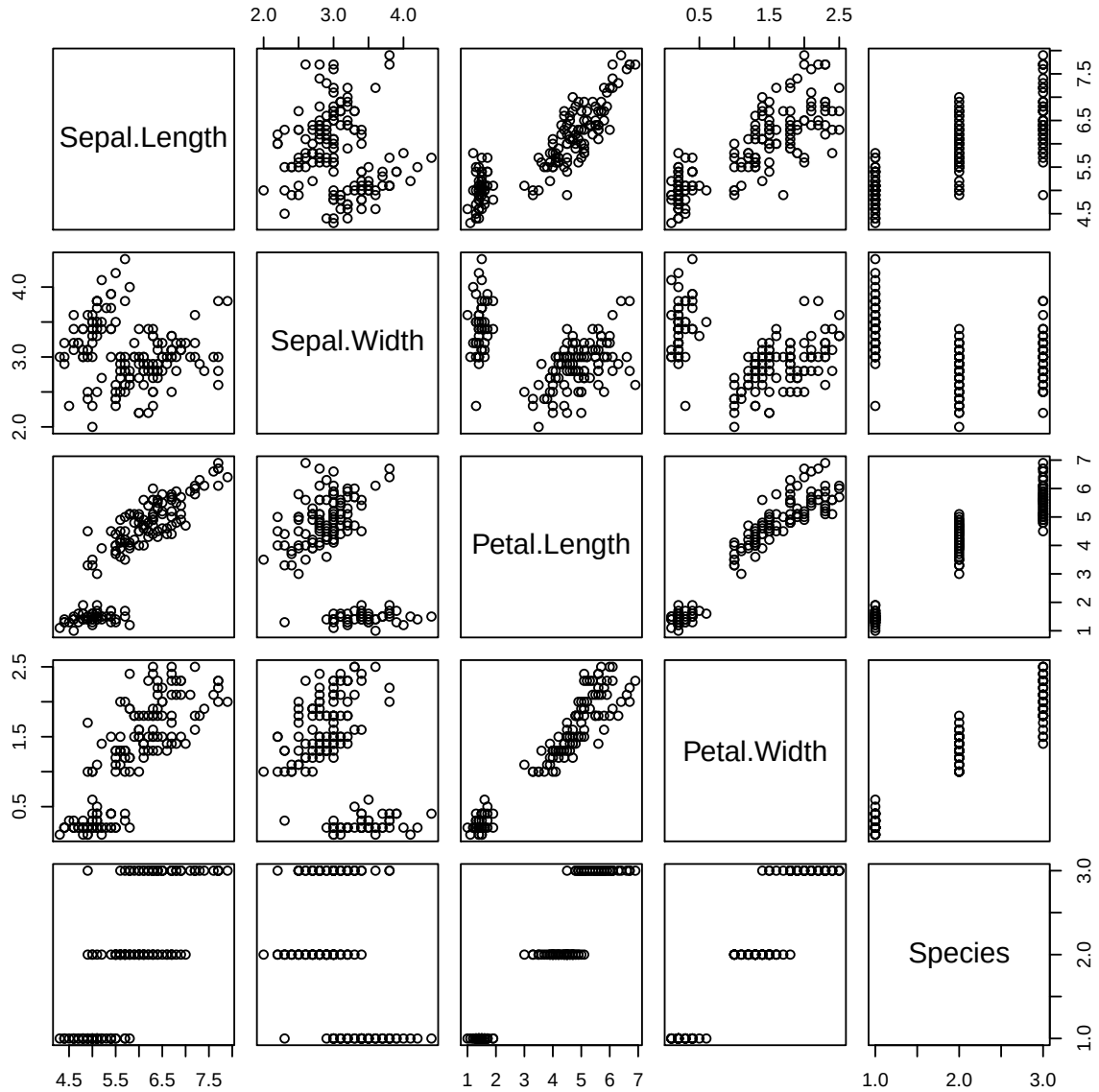
Three Variables

Two Quantitative Variables and One Qualitative Variable

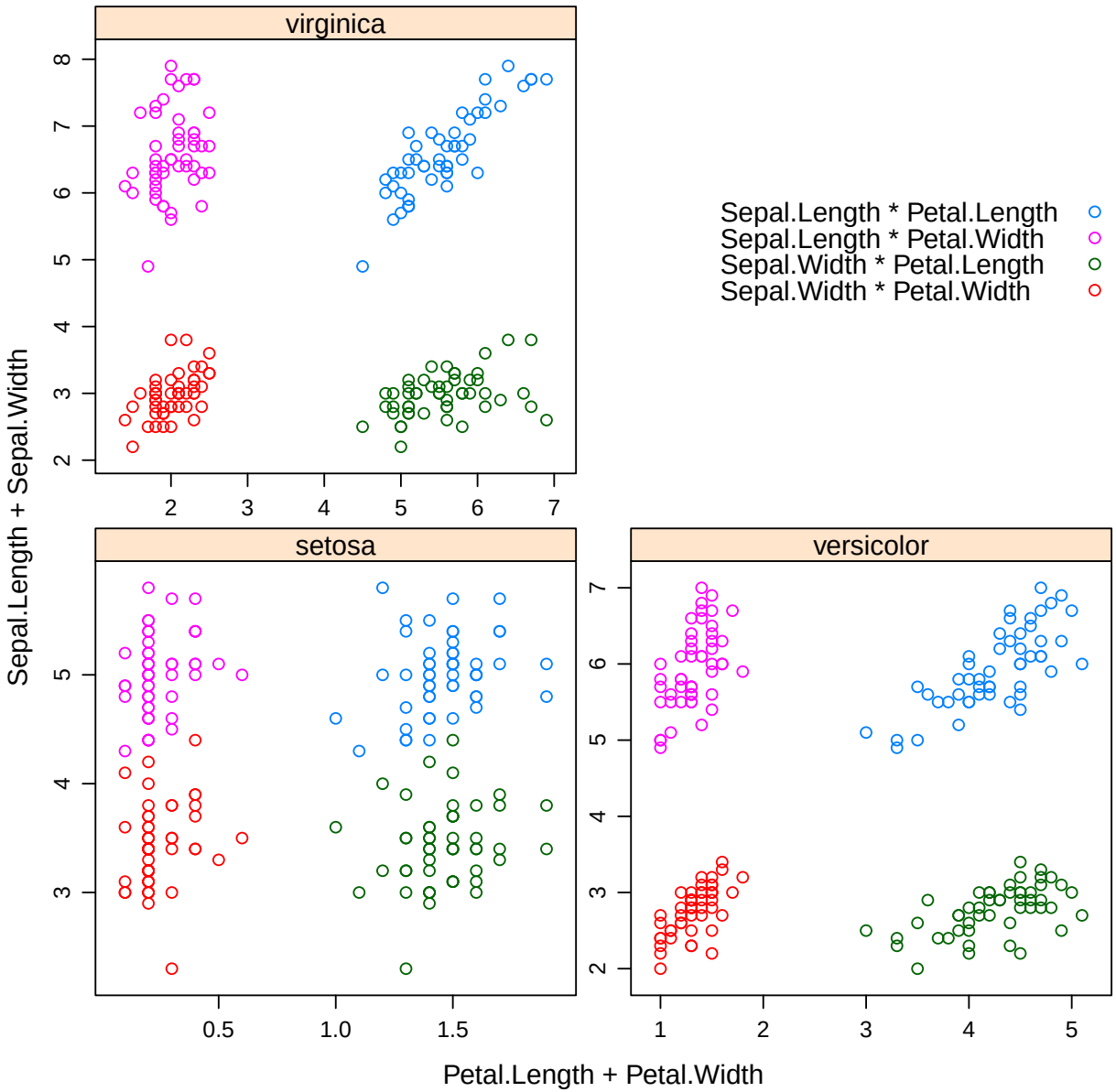


More Than Three Variables

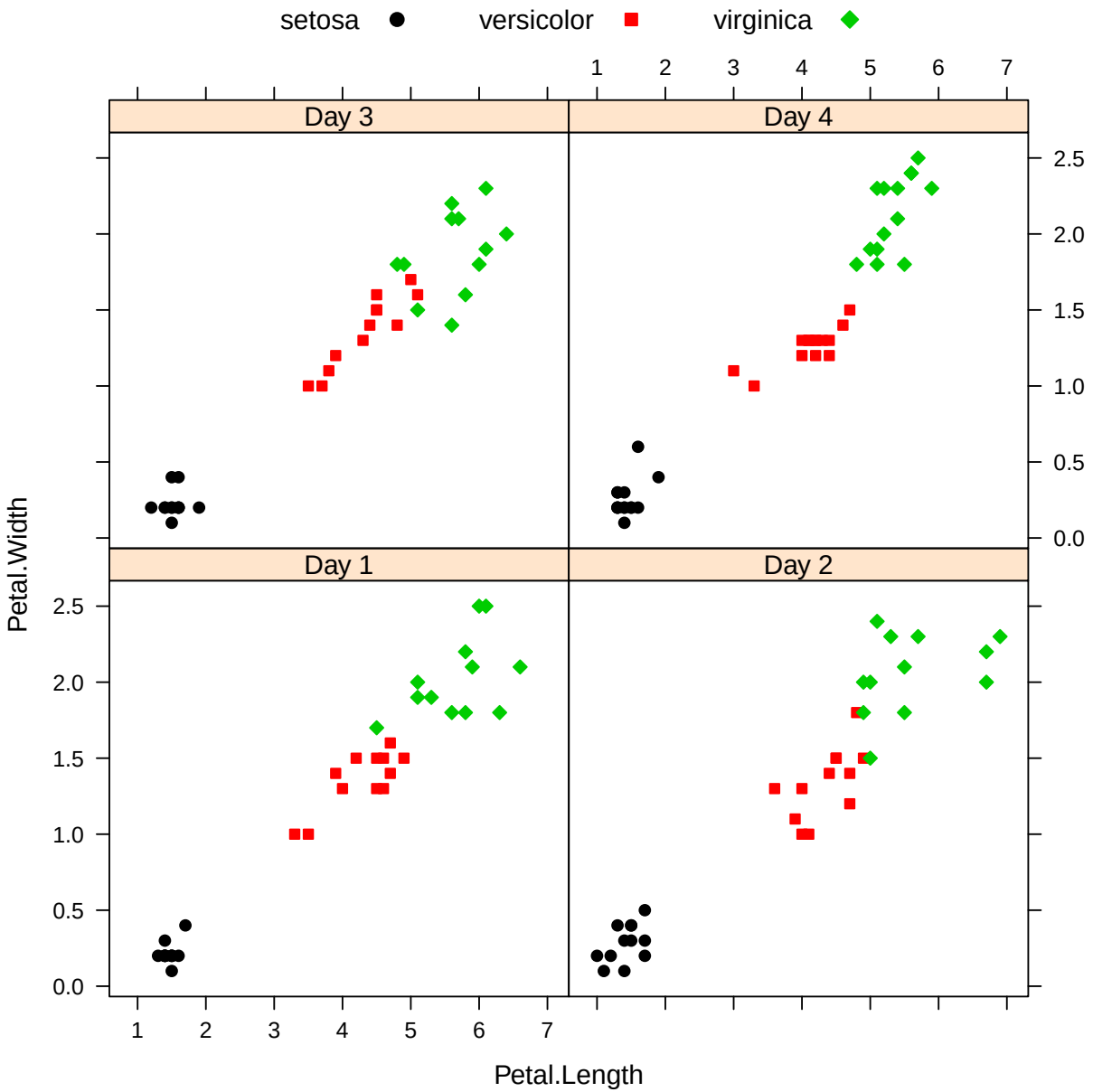
Pairs Plot



More Than Three Variables

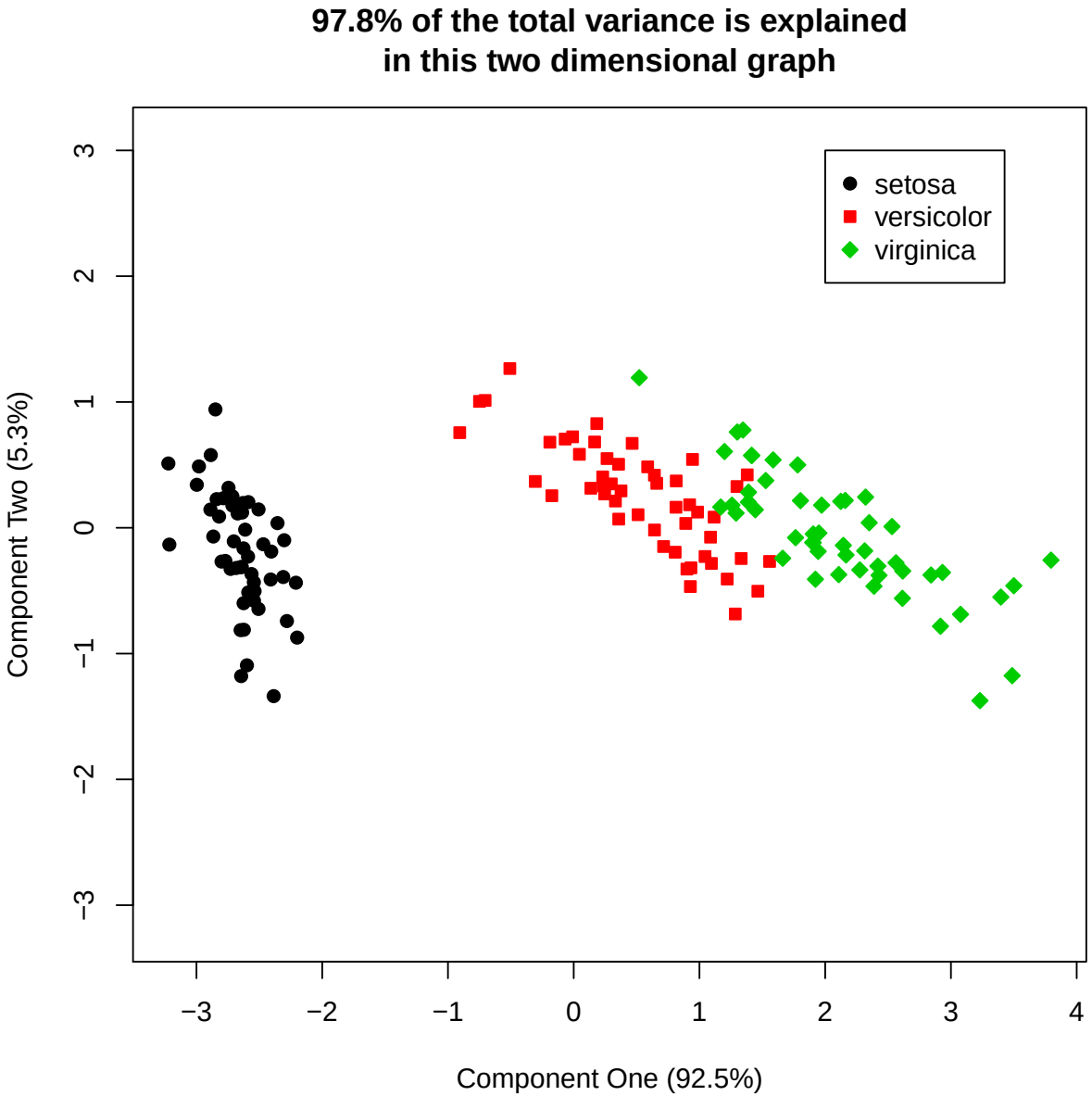


More Than Three Variables



Multivariate

- Suppose a dataset has n observations (rows) and p variables (columns).
- The n observations lie in a p dimensional space.
- To visualise the data, we apply some mathematical procedures to reduce the dimensionality of the data cloud, optimally into two dimensions.
  1. Throw away some/lots of variances/information, TANSTAAFL.
  2. Plot the reduced data into a two/three dimensional graphs, and visualise the MAIN components of variance.


97.8% of the total variance is explained
in this two dimensional graph
setosa
versicolor
virginica
Component Two (5.3%)
Component One (92.5%)
−3
−2
−1
0
1
2
3
4

Recommended Reading

Kai.

`http://www.stat.auckland.ac.nz/~kxio001`.

Paul Murrell.

R Graphics.

Chapman & Hall/CRC, Boca Raton, FL, 2005.