Data Analytics
SET10109

Understanding Data Using Visualisation

Lecturer: David Hunter

Course Content: Natalie Kerracher and David Hunter (d.hunter@napier.ac.uk)

Lecture plan

This lecture looks at how we can use visual representations to explore the properties of our data set. We will cover:

- Data types
- Visual Encodings
- Some basic ways that you could plot data to gain an initial understanding of your dataset

Reading

REQUIRED:

- Chapter 4: Data Understanding in Berthold, Borgelt, Höppner, and Klawonn. Guide to intelligent data analysis. Vol. 42. Springer, 2010.

PLEASE MAKE SURE THAT YOU READ THIS CHAPTER!!

(NB Taoxin will cover data quality issues in the next lecture)

Recommended

- Heer, J., Bostock, M., & Ogievetsky, V. (2010). A tour through the visualization zoo. *Commun. ACM*, *53*(6), 59-67. Available at http://queue.acm.org/detail.cfm?id=1805128

VISUAL PROPERTIES – A TASTER

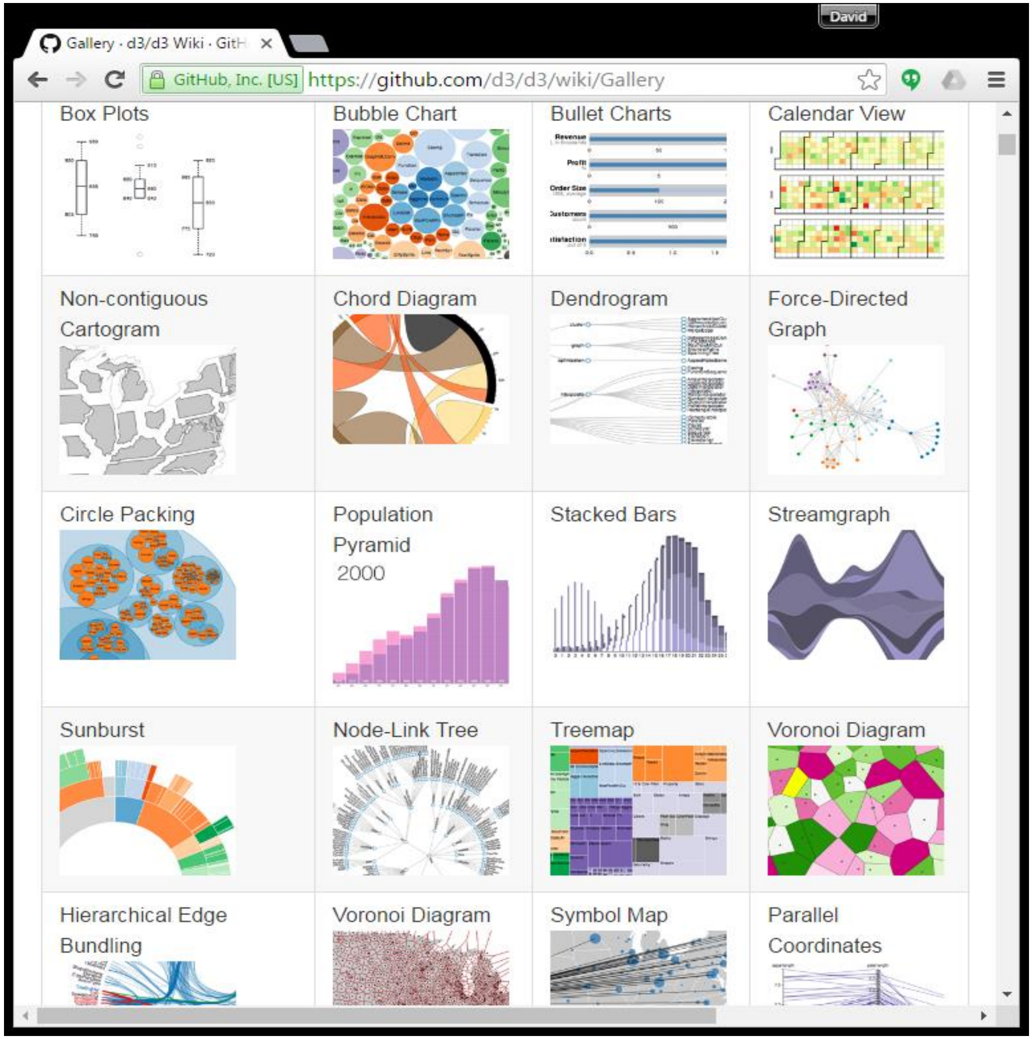
Visual encoding – what is it?

By visual encoding, we simply mean mapping a data attribute to a visual variable.

Visualisation depends on:

Type of data being viewed

The questions we want to ask.

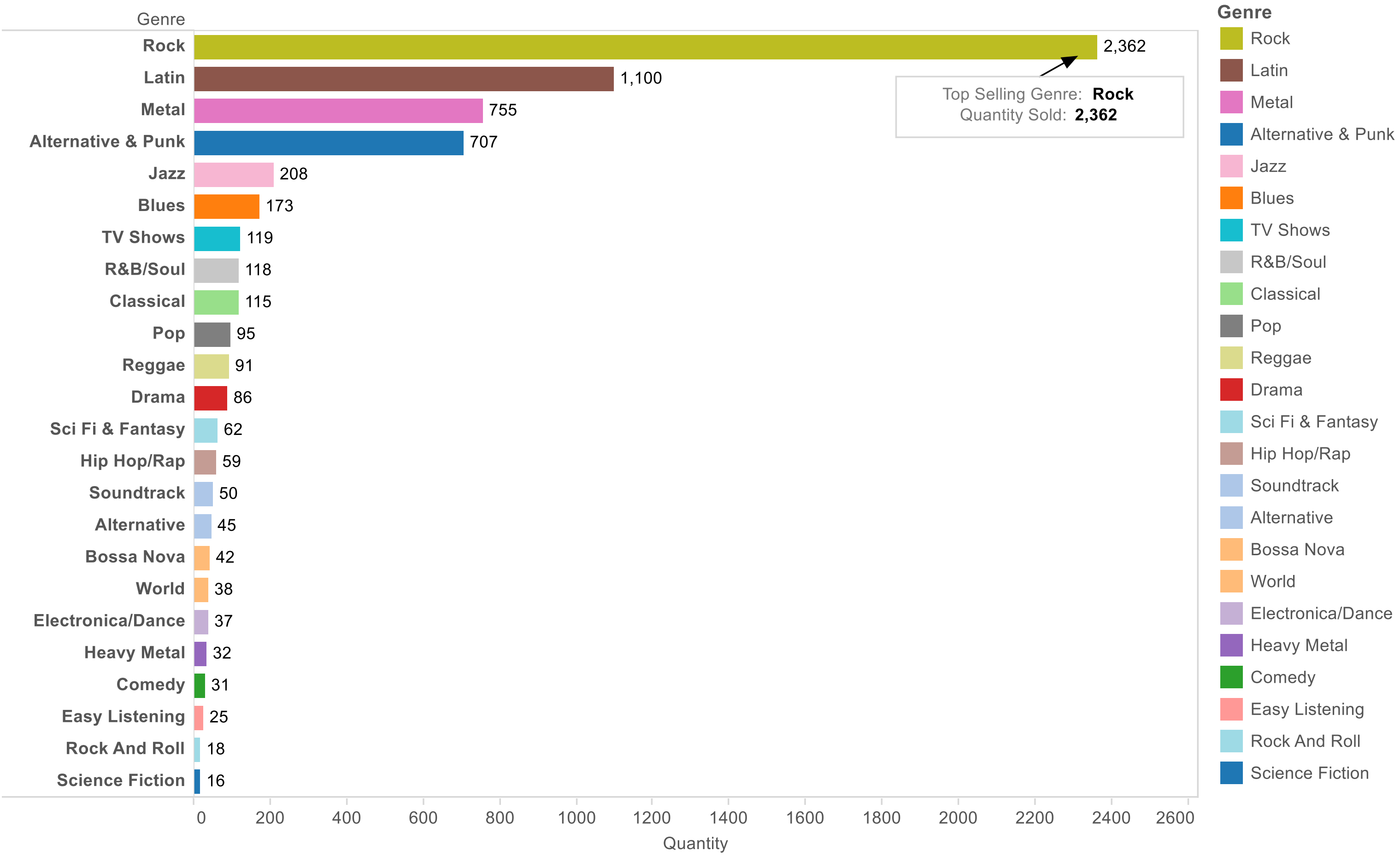


https://github.com/d3/d3/wiki/Gallery

Some plot types

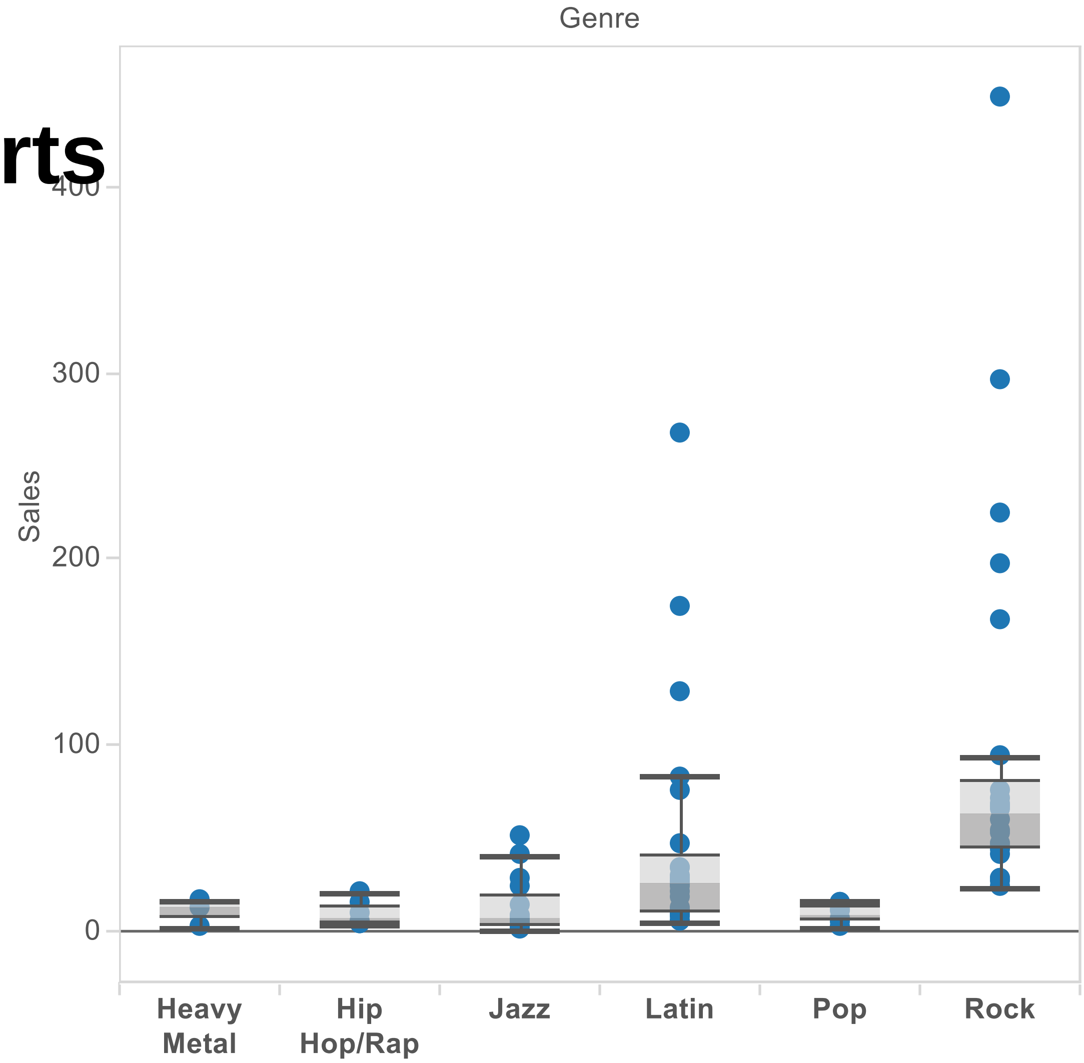
- Bart charts
- Box charts

Bar charts

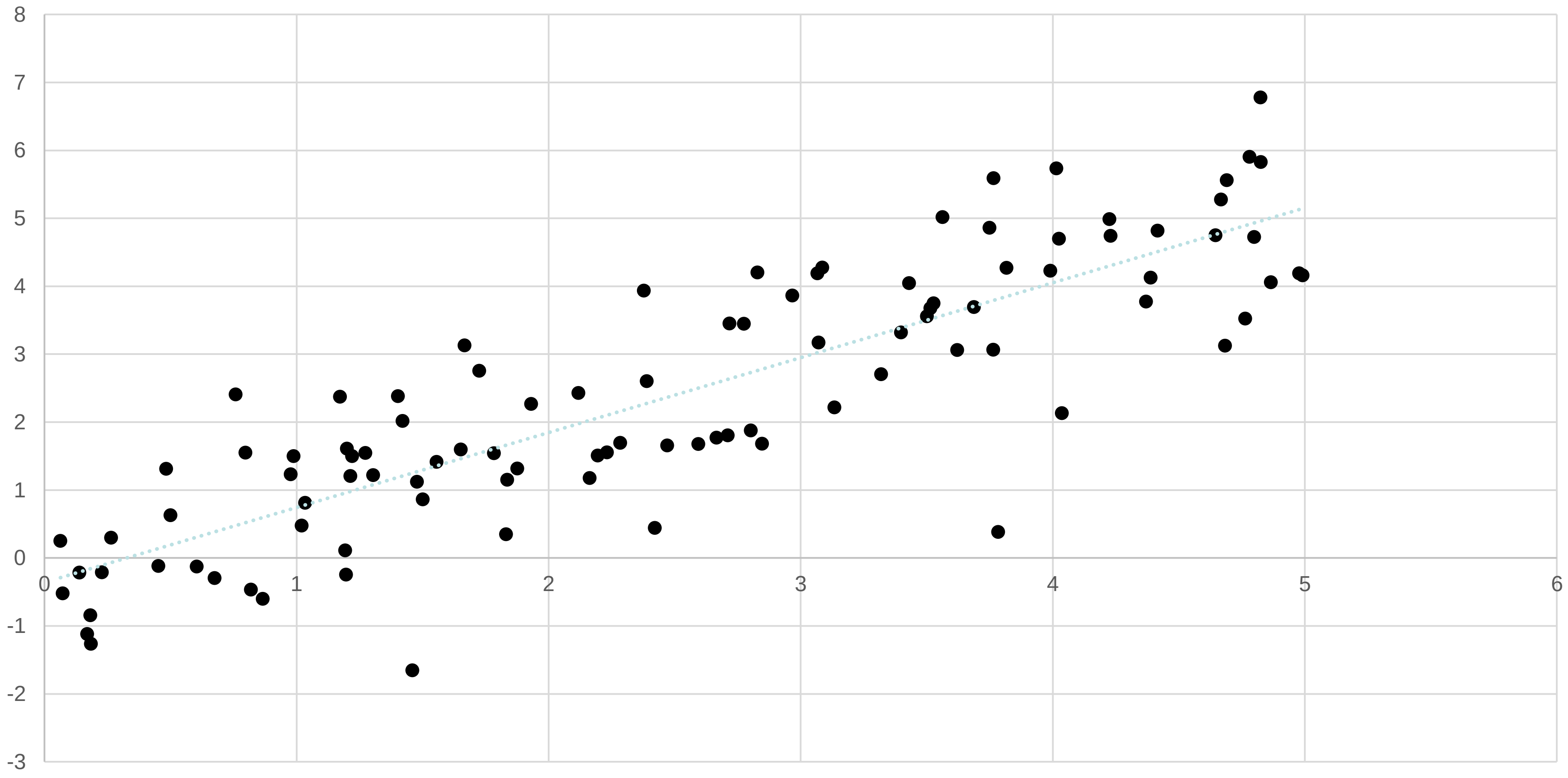


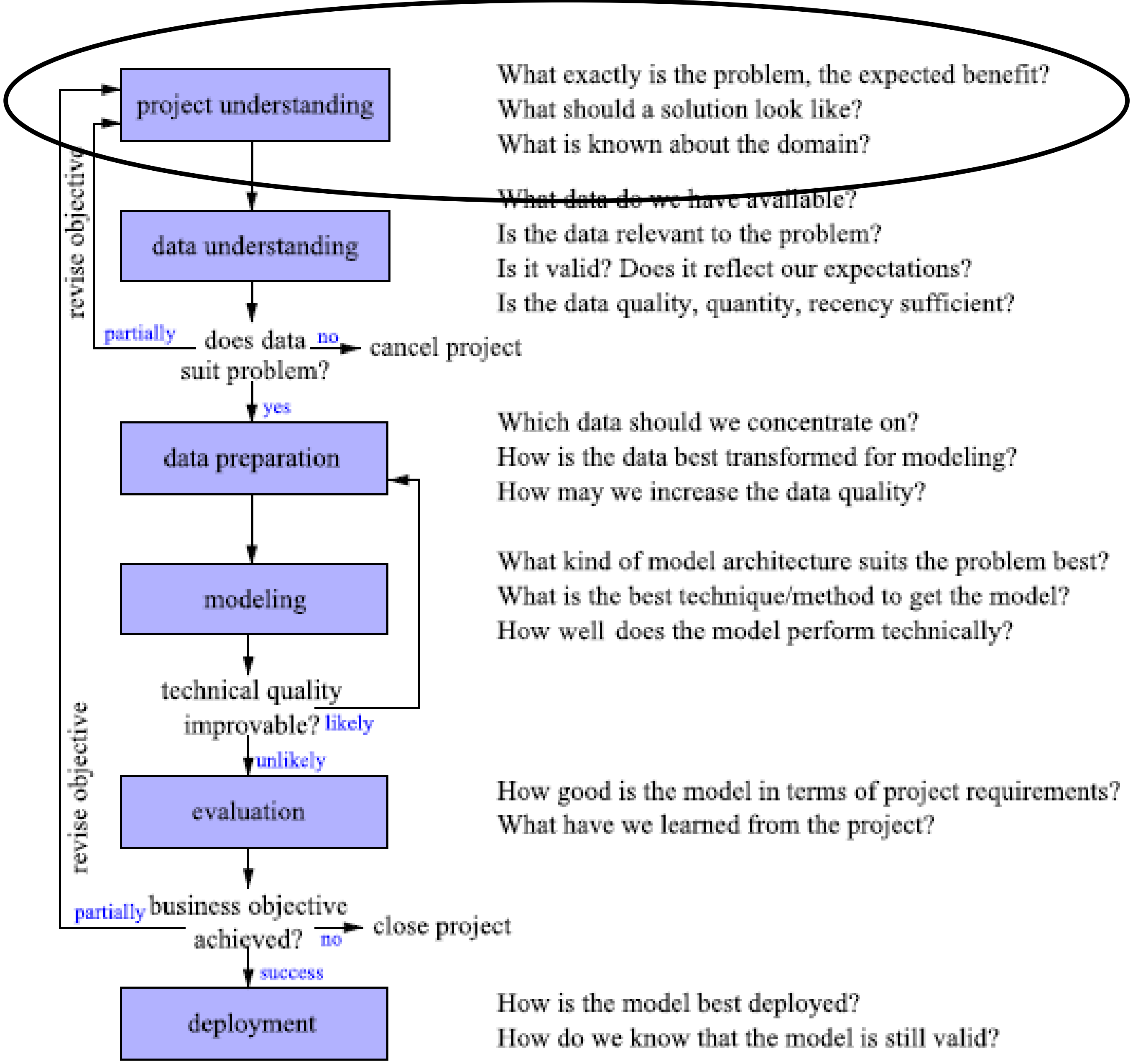
Top Selling Genre: **Rock**
Quantity Sold: **2,362**

Box charts


Genre
Sales
400
300
200
100
0
Heavy Metal
Hip Hop/Rap
Jazz
Latin
Pop
Rock

Scatter plot





What exactly is the problem, the expected benefit?
What should a solution look like?
What is known about the domain?

What data do we have available?
Is the data relevant to the problem?
Is it valid? Does it reflect our expectations?
Is the data quality, quantity, recency sufficient?

Which data should we concentrate on?
How is the data best transformed for modeling?
How may we increase the data quality?

What kind of model architecture suits the problem best?
What is the best technique/method to get the model?
How well does the model perform technically?

How good is the model in terms of project requirements?
What have we learned from the project?

How is the model best deployed?
How do we know that the model is still valid?

Fig. 1.1 Overview of the CRISP-DM process together with typical questions to be asked in the respective phases

Figure from Berthold et al. (2010), p9

Taking a step back (Project Understanding)

Context. You need to know in advance..

What is the background of the project?

What are the overall aims of the project?

What restrictions are placed on the project?

Taking a step back (Project Understanding)

Domain. You need to know in advance..

What do we know about the problem?

What sort of effects should be look for?

What prior expectations we have?

Taking a step back (Project Understanding)

Audience. You need to know in advance..

Who will view the visualisation?

What expectations do the viewers have?

What message you expect to get across?

Taking a step back (Project Understanding)

Solution. You need to think in advance..

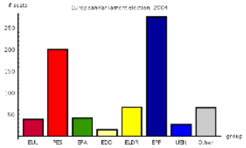
What do you expect the solution to look like.

Types of Visualisation

The choice of Visualization depends on what we want to learn from it:

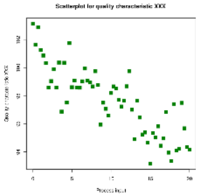
The distribution of data:

1 variable charts (bar, histogram..)



Interactions between variables

2 or more variable charts (scatterplot etc.)



Planning

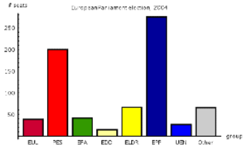
e.g. maps

Types of Visualisation

The choice of Visualization depends on what we want to learn from it:

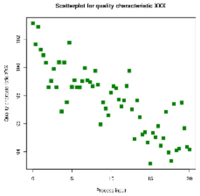
The distribution of data:

1 variable charts (bar, histogram..)



Interactions between variables

2 or more variable charts (scatterplot etc.)



Planning

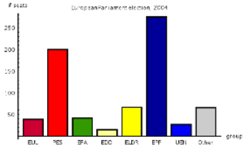
e.g. maps

Types of Visualisation

The choice of Visualization depends on what we want to learn from it:

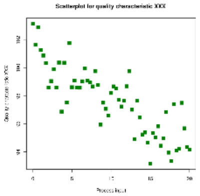
The distribution of data:

1 variable charts (bar, histogram..)



Interactions between variables

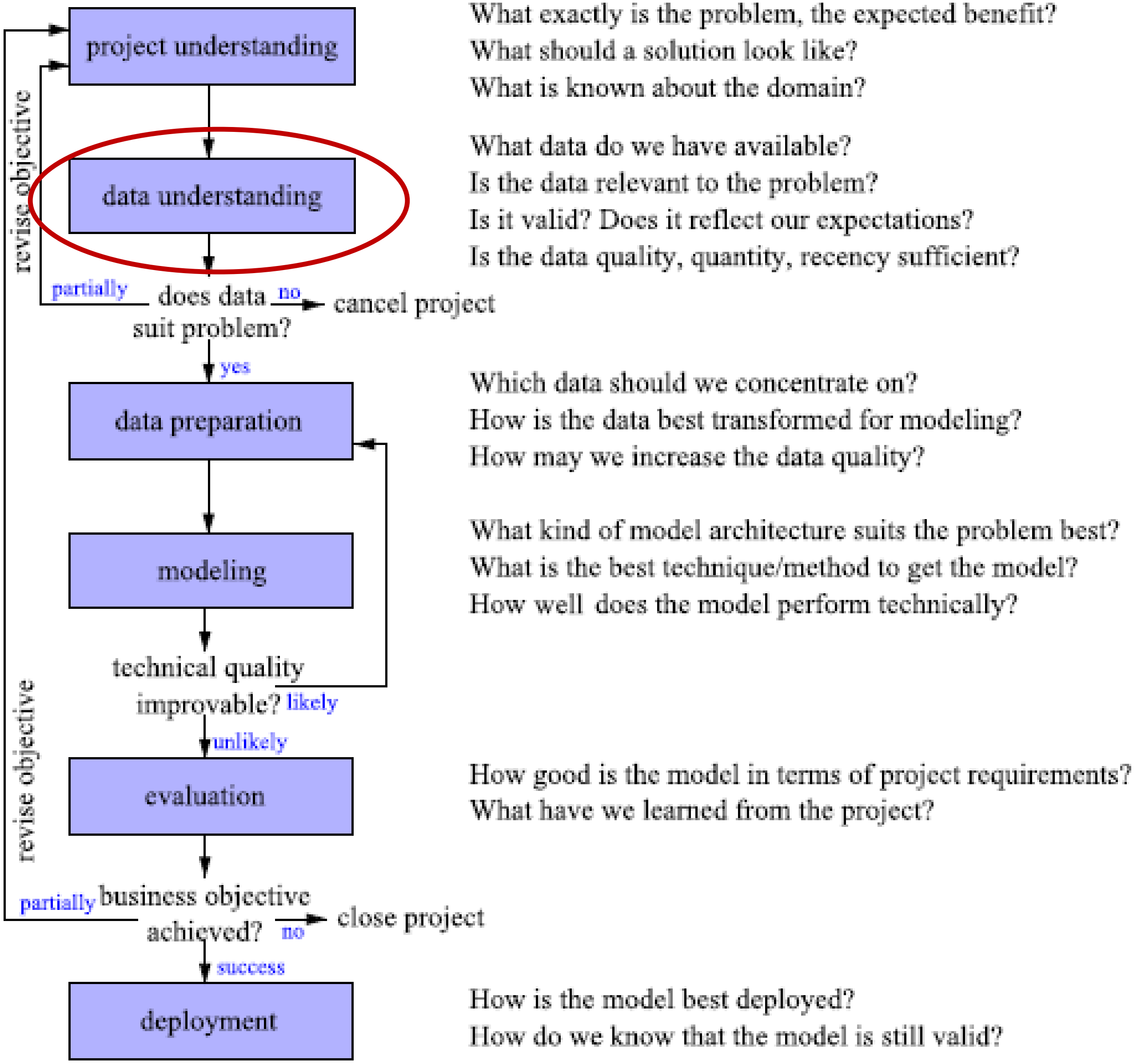
2 or more variable charts (scatterplot etc.)



Planning

e.g. maps

We are here!



What exactly is the problem, the expected benefit?
What should a solution look like?
What is known about the domain?

What data do we have available?
Is the data relevant to the problem?
Is it valid? Does it reflect our expectations?
Is the data quality, quantity, recency sufficient?

Which data should we concentrate on?
How is the data best transformed for modeling?
How may we increase the data quality?

What kind of model architecture suits the problem best?
What is the best technique/method to get the model?
How well does the model perform technically?

How good is the model in terms of project requirements?
What have we learned from the project?

How is the model best deployed?
How do we know that the model is still valid?

Fig. 1.1 Overview of the CRISP-DM process together with typical questions to be asked in the respective phases

Figure from Berthold et al. (2010), p9

Data understanding

Process (and lesson plan)

1. What sort of data to we have?
   - Numbers, locations, categories?
   - How many and what sort of **interactions** do we expect.

2. Does the data meet our expectations?
   - What range and distribution do we expect.
   - Use visualisation to check our assumptions.

Look at the raw data

What sort of data are we dealing with? Numbers, Categories, Locations?

Tabular: Recorded in a table e.g. EXCEL, SPSS, Tableau

Continuous (allows fraction) vs Discrete (no factions)

Number of dimensions (normally columns in a table)

Size (number of records/data points)

Visually identify input errors.

Does the data look like what you expected

Attribute understanding

In this lecture, we will assume that the data set is provided in the form of a simple table (tabular data)

	attribute$_1$	**…**	**attribute$_m$**
record$_1$			
: :			
record$_n$			

- The rows of the table are called instances, records or data objects.
- The columns of the table are called attributes, features or variables.

Slide based on Compendium slides for Guide to Intelligent Data Analysis, Springer 2011. c Michael R. Berthold, Christian Borgelt, Frank H•oppner, Frank Klawonn and Iris Ad (slide 3)

Type of attributes

Discrete: (Qualitative) Values that cannot be placed on a range

Green, Blue, Brown

Male, Female

Continuous: (Quantitative) Values that can define a difference or separation + an order (larger/smaller) i.e. a measure.

Meters

Temperature

Type of attributes

Categorical: Discrete categories.

Green, Blue, Brown

Male, Female.

Ordinal: Ordered sets of values.

Small, Medium, Large

1st, 2nd, 3rd.

Interval: Ordered set of values with known separation (a measurement)

Temperature in Celsius

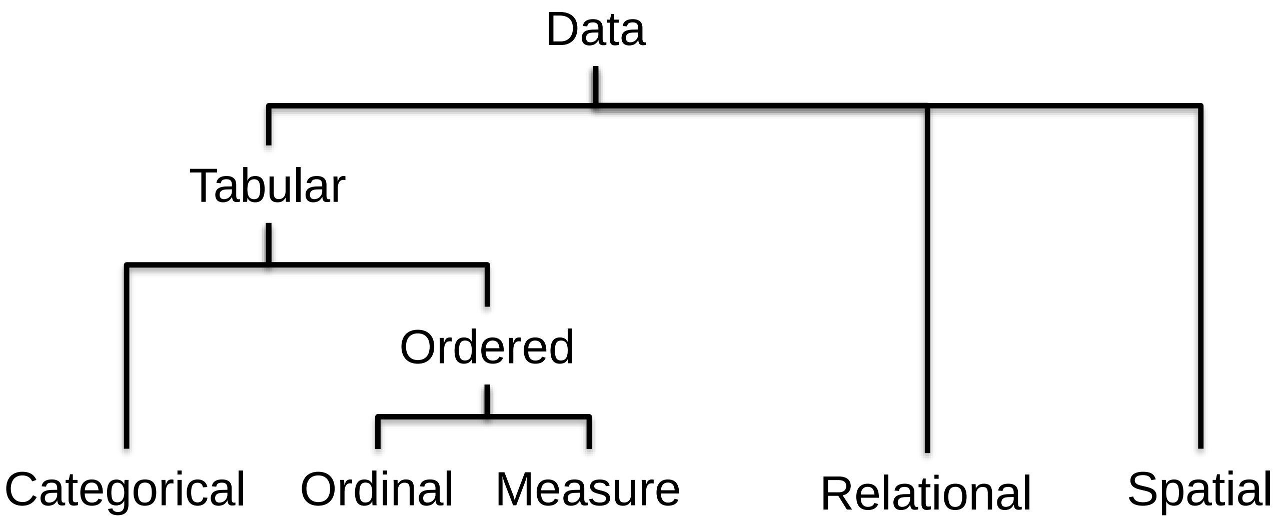
Ratio: Interval + meaningful zero.

Height (m)

Weight (Kg)

Temperature (Kelvin)

Types of attribute



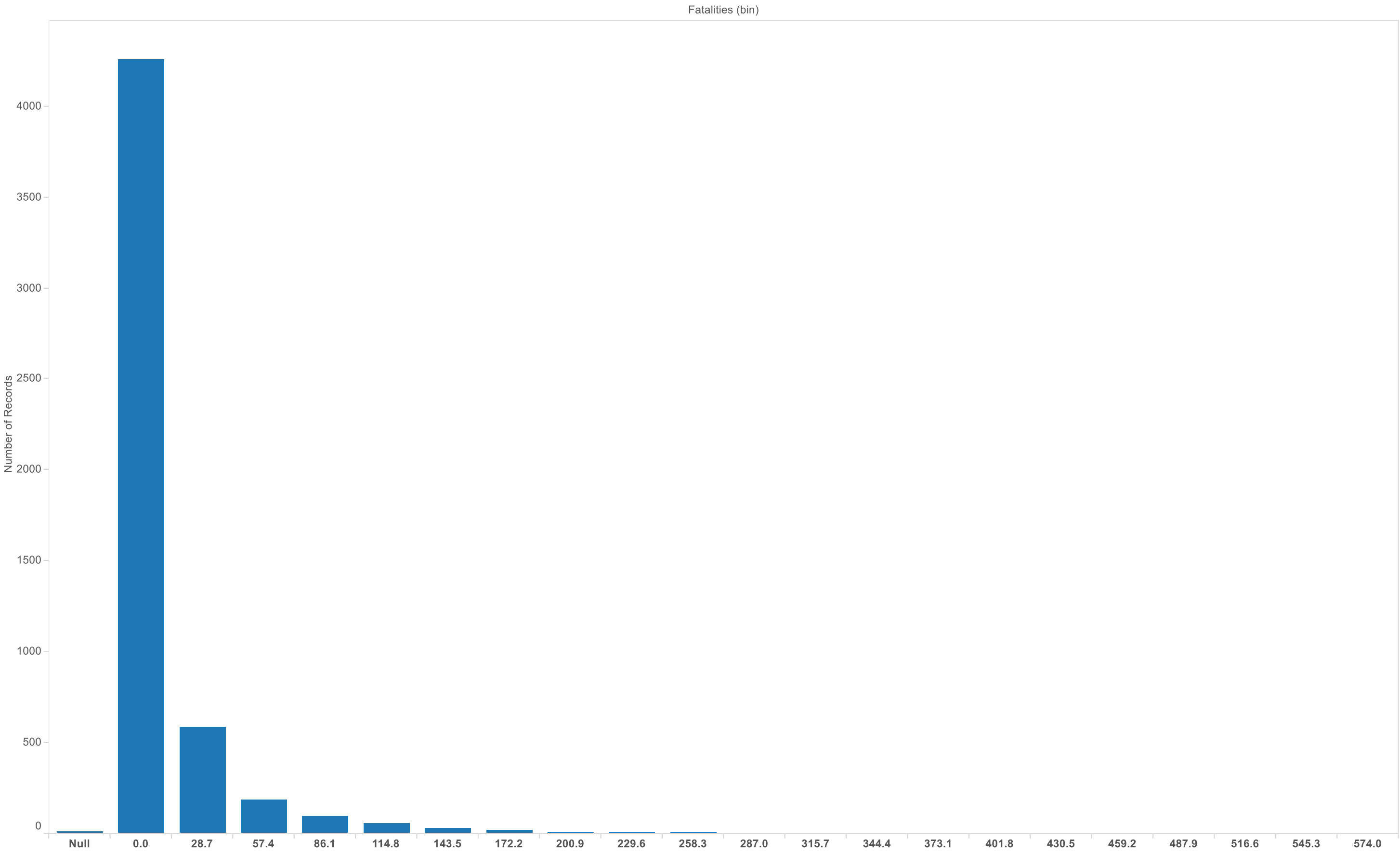
Data Exploration

First stage of a data analysis

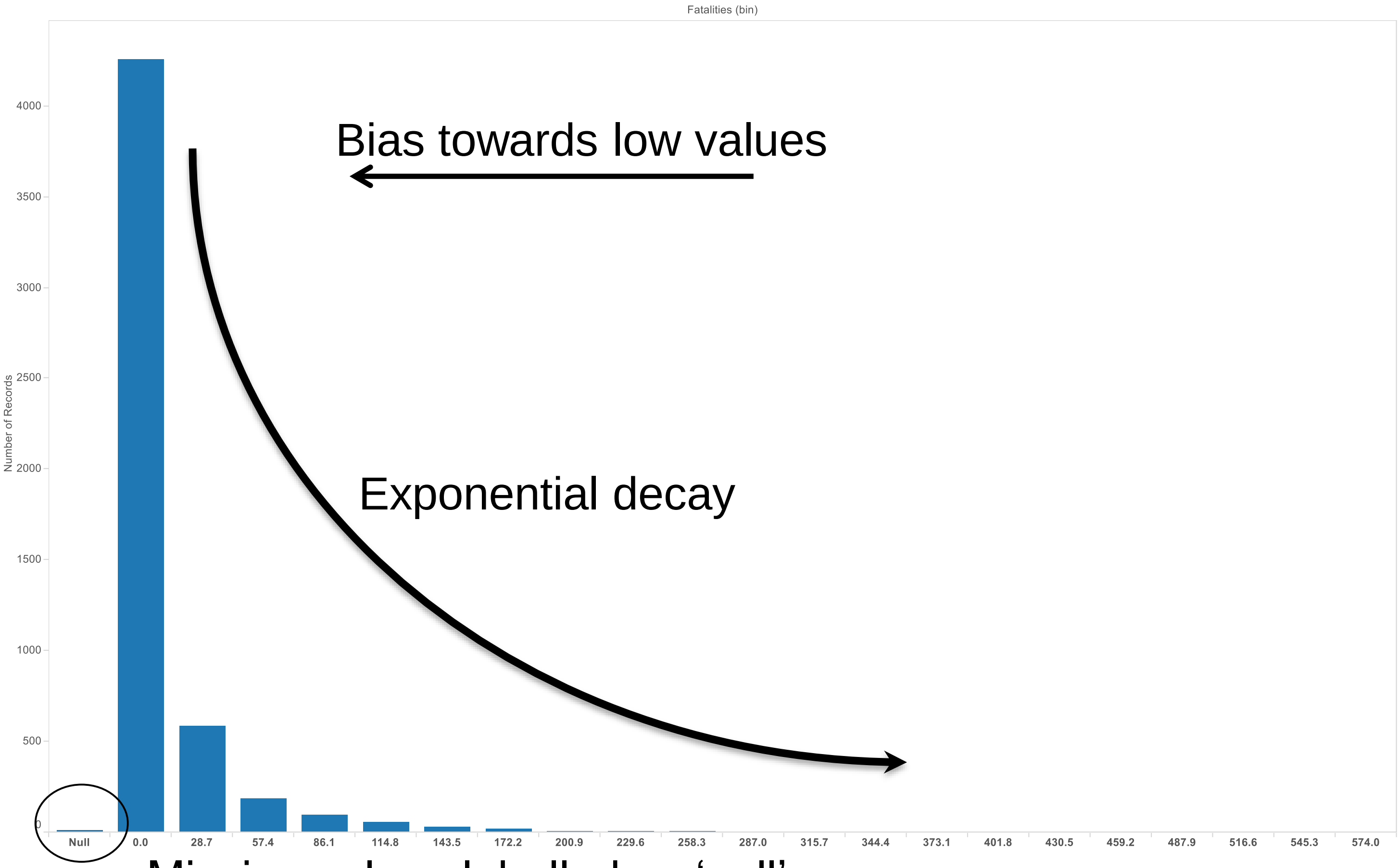
Aims

- Understand the dataset
  - What is the range and limitations of the data?
- Check the data
  - Is the data consistent with expectations?
- Understand relationships in the dataset
  - What links (if any) exist between items in the dataset

Analysing single variable



Analysing single variable



Bias towards low values

Exponential decay

Missing values labelled as ‘null’

Analysing single variable

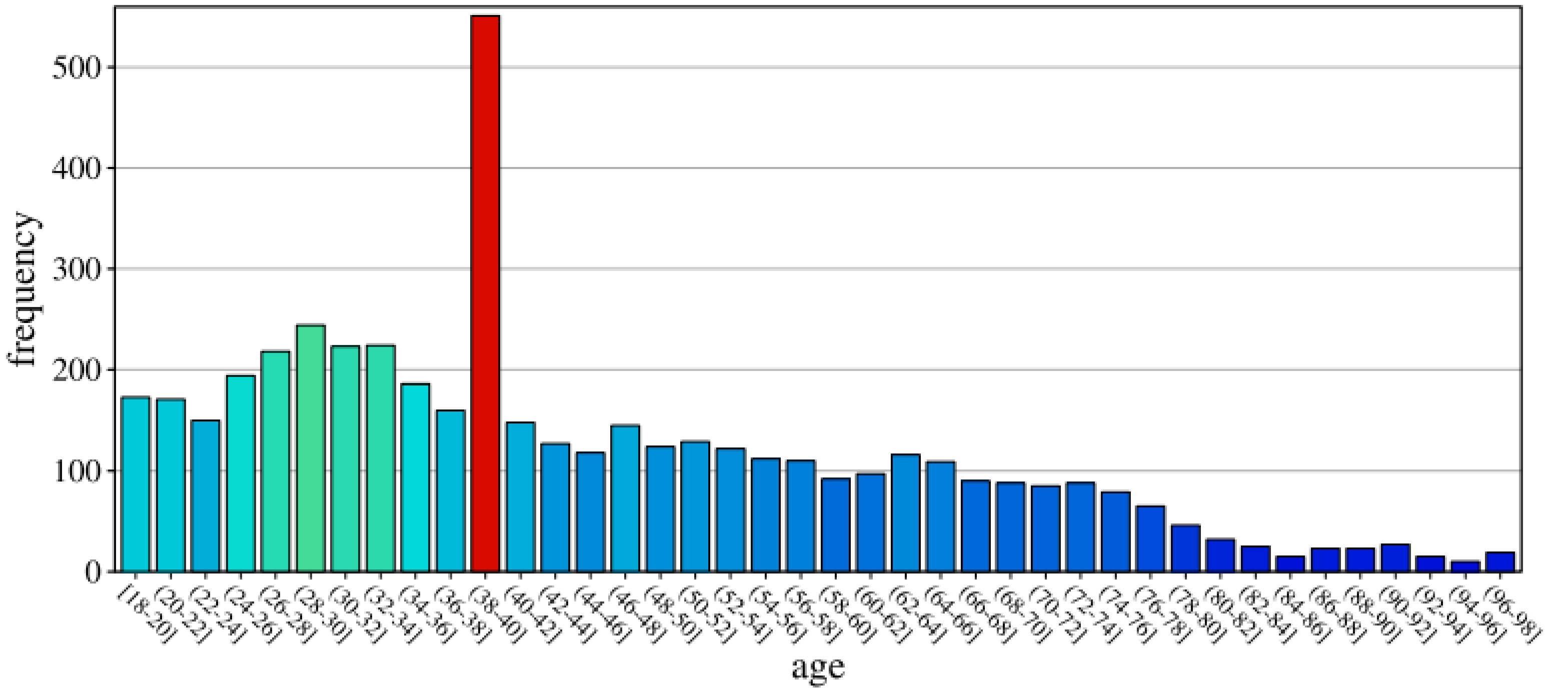
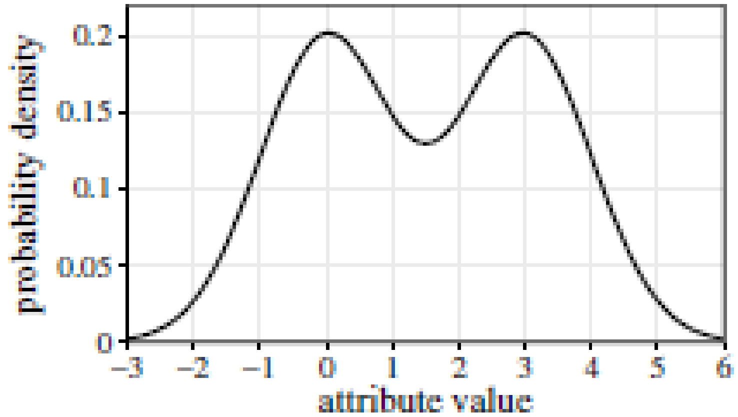
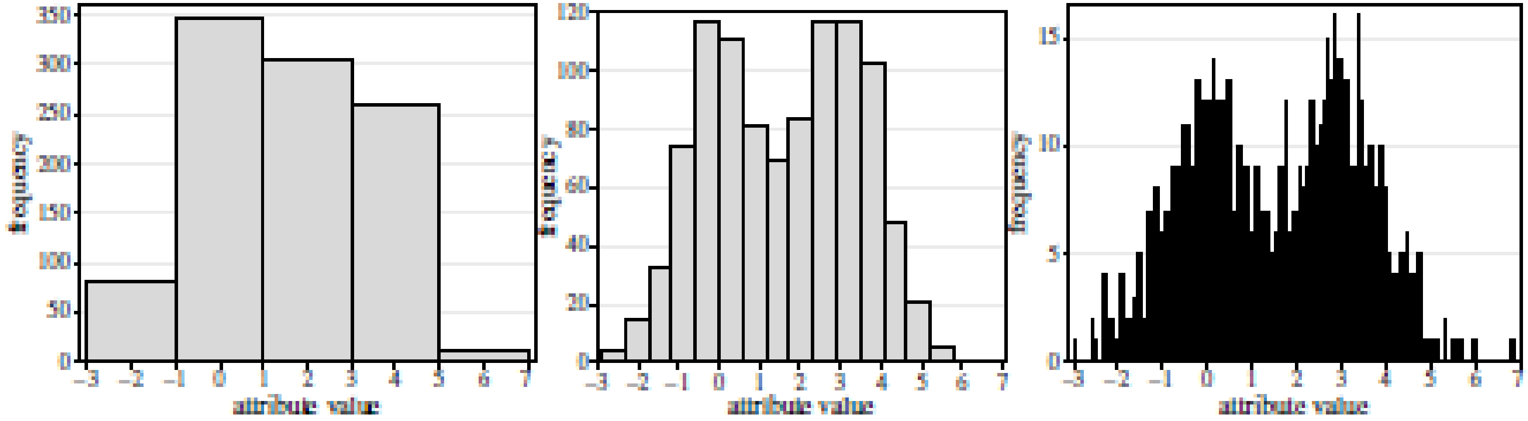


Fig. 2.2 A histogram for the distribution of the value of attribute *age* using 40 bins

Histograms: Number of bins





Three histograms with 5, 17 and 200 bins for a sample from the same bimodal distribution.

Distributions

Distribution

“the way in which something is shared out among a group or spread over an area.” – google.com

A distribution describes rate of occurrence of values within a variable.

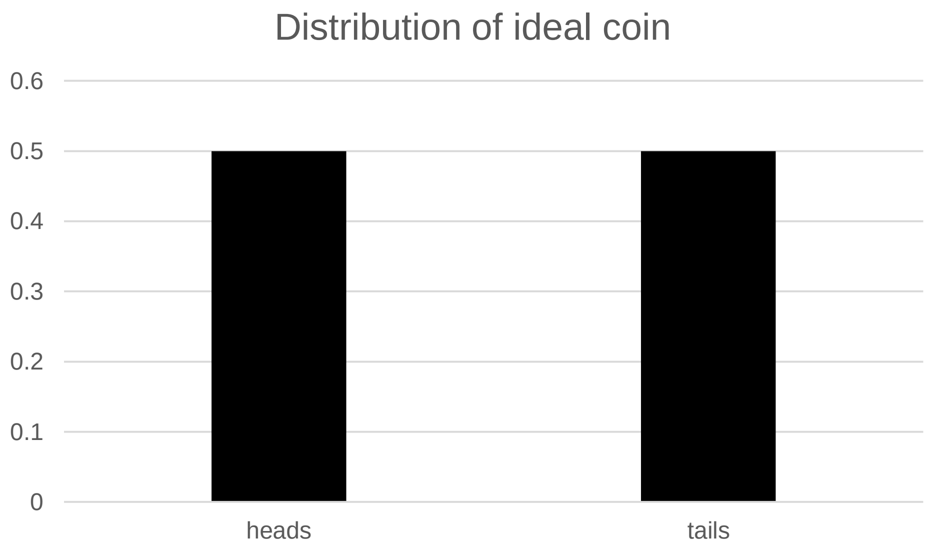
Distributions

Examples

A coin toss (category). For a balanced coin we would expect half heads and half tails.

0.5 – head

0.5 – tails.



Distributions

Examples

A 6 sided die (interval)

1 – 1/6

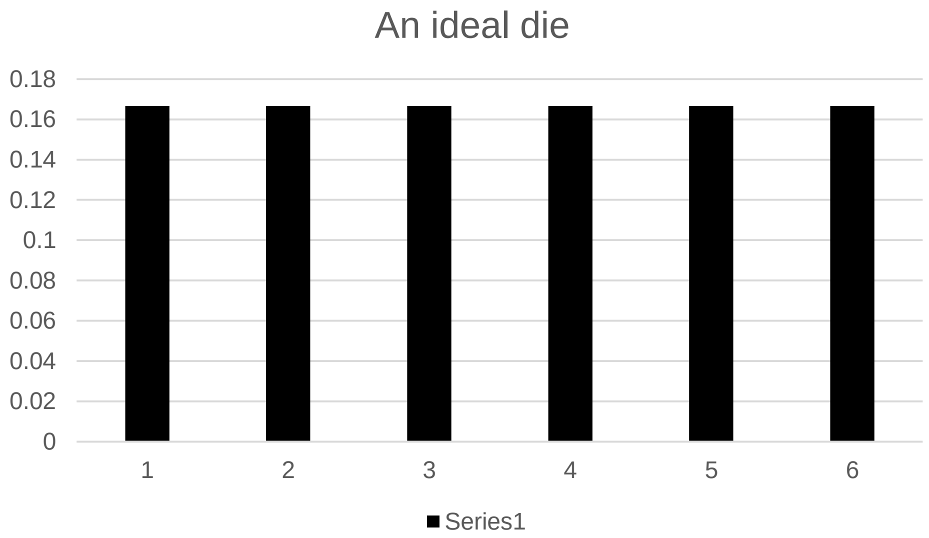
2 – 1/6

3 – 1/6

4 – 1/6

5 – 1/6

6 – 1/6



Distributions

Examples

A 6 sided die (interval) – unfair die

1 – 1/12

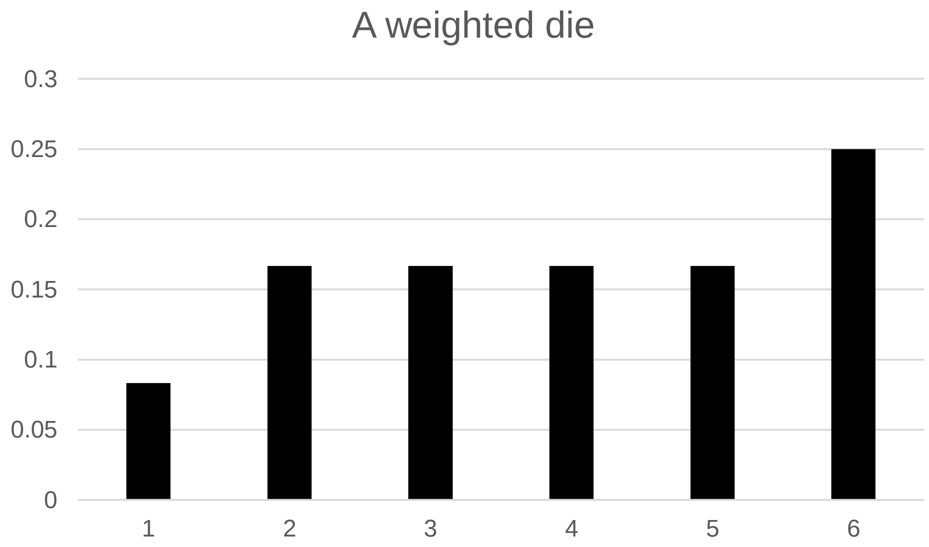
2 – 1/6

3 – 1/6

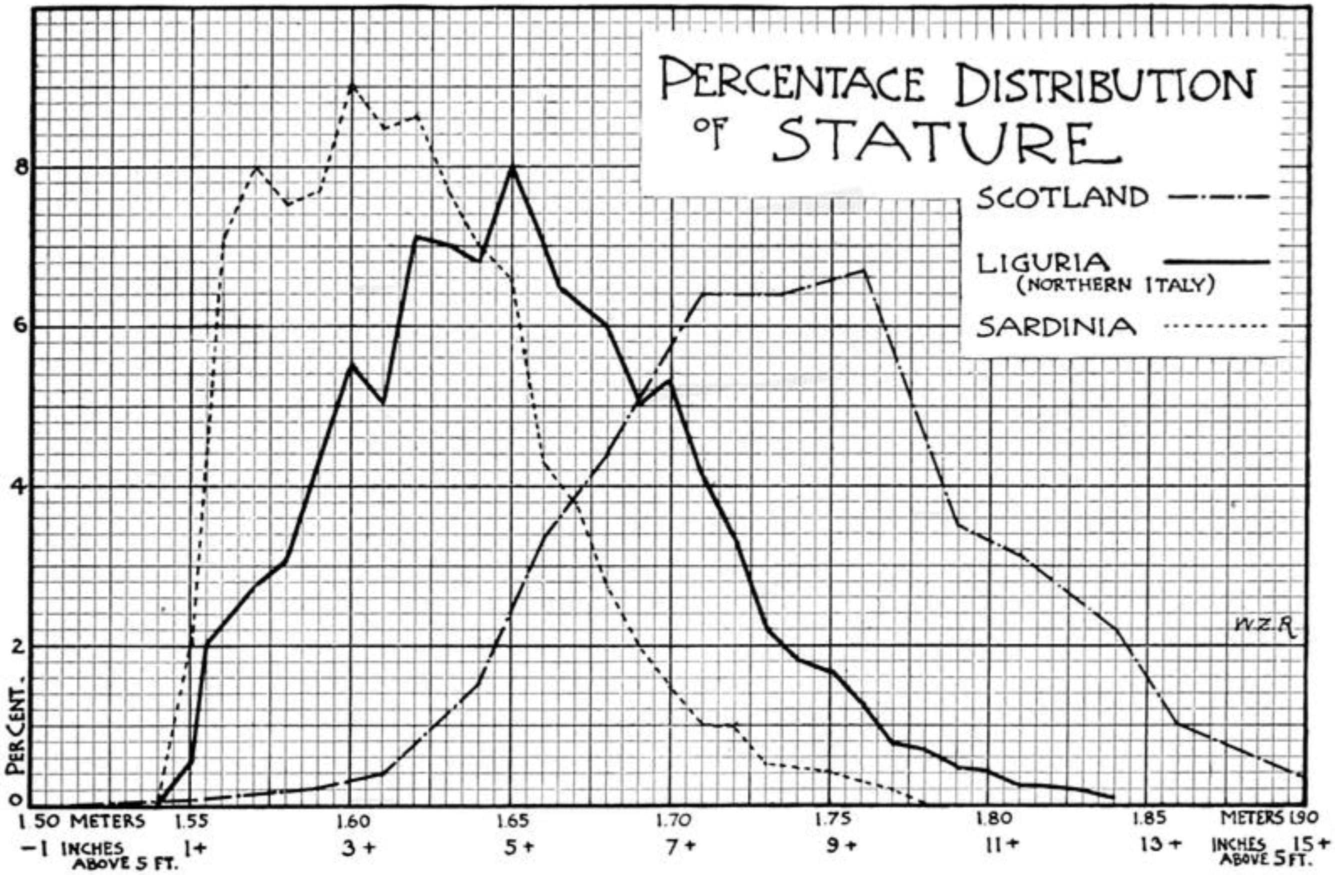
4 – 1/6

5 – 1/6

6 – 3/12



Distributions



Distributions

The distribution is defined over all possible values of the variable (including non-observed values).

Distributions can be plotted as:

Histograms – measures, intervals, ratios

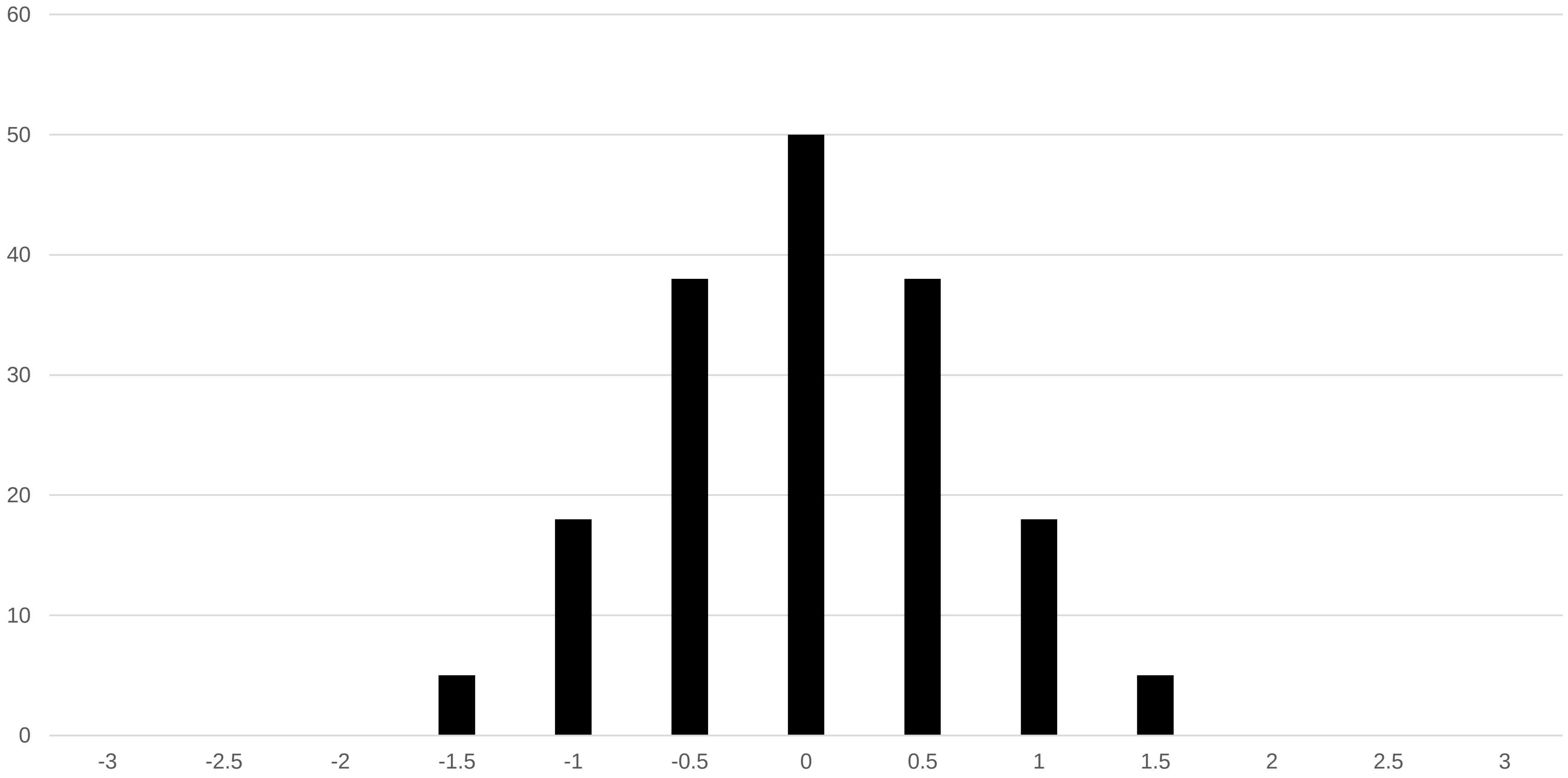
Bar charts – categories

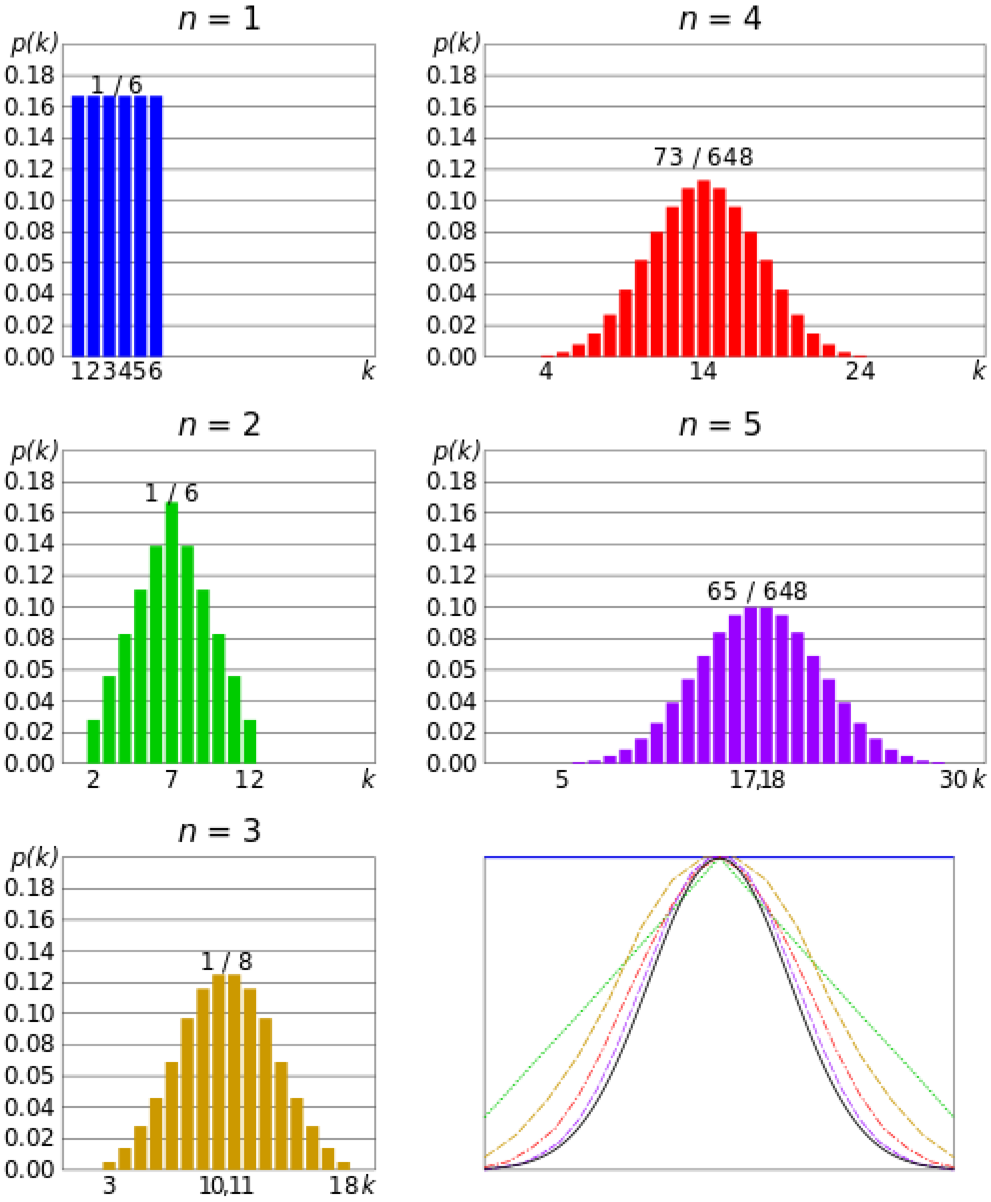
All probabilities must sum to 1.

Histograms do not have to sum to 1 but they must sum to the number of samples.

Convert from histogram to distribution by: dividing every bin count by the total count for all samples.

Normal Distribution





Example of distributions visualised with a Histogram.

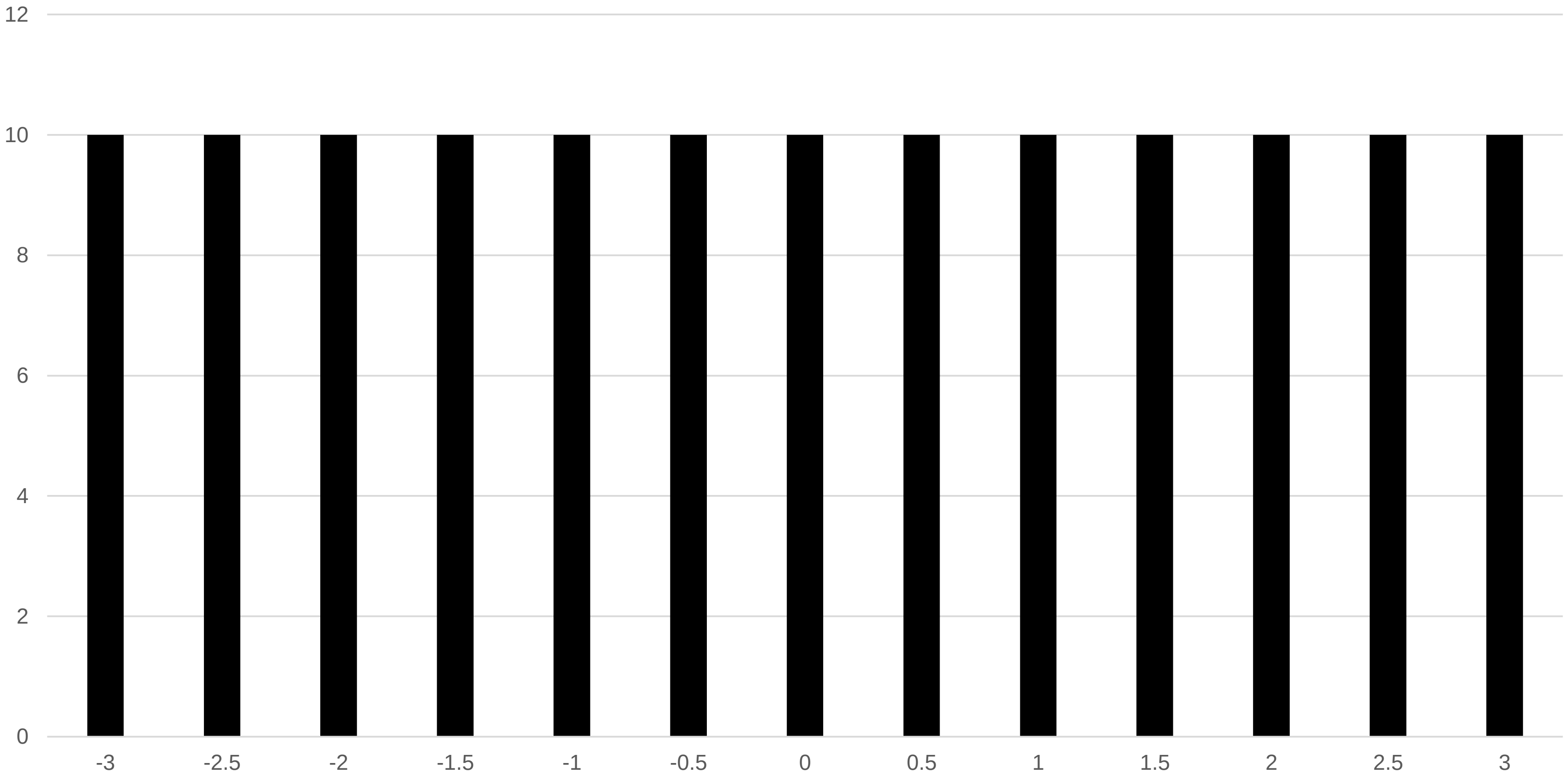
Top left: Results of a single fair die. Uniform distribution

Top right: Sum of 4 dice. Beginning to approximate a Normal distribution.

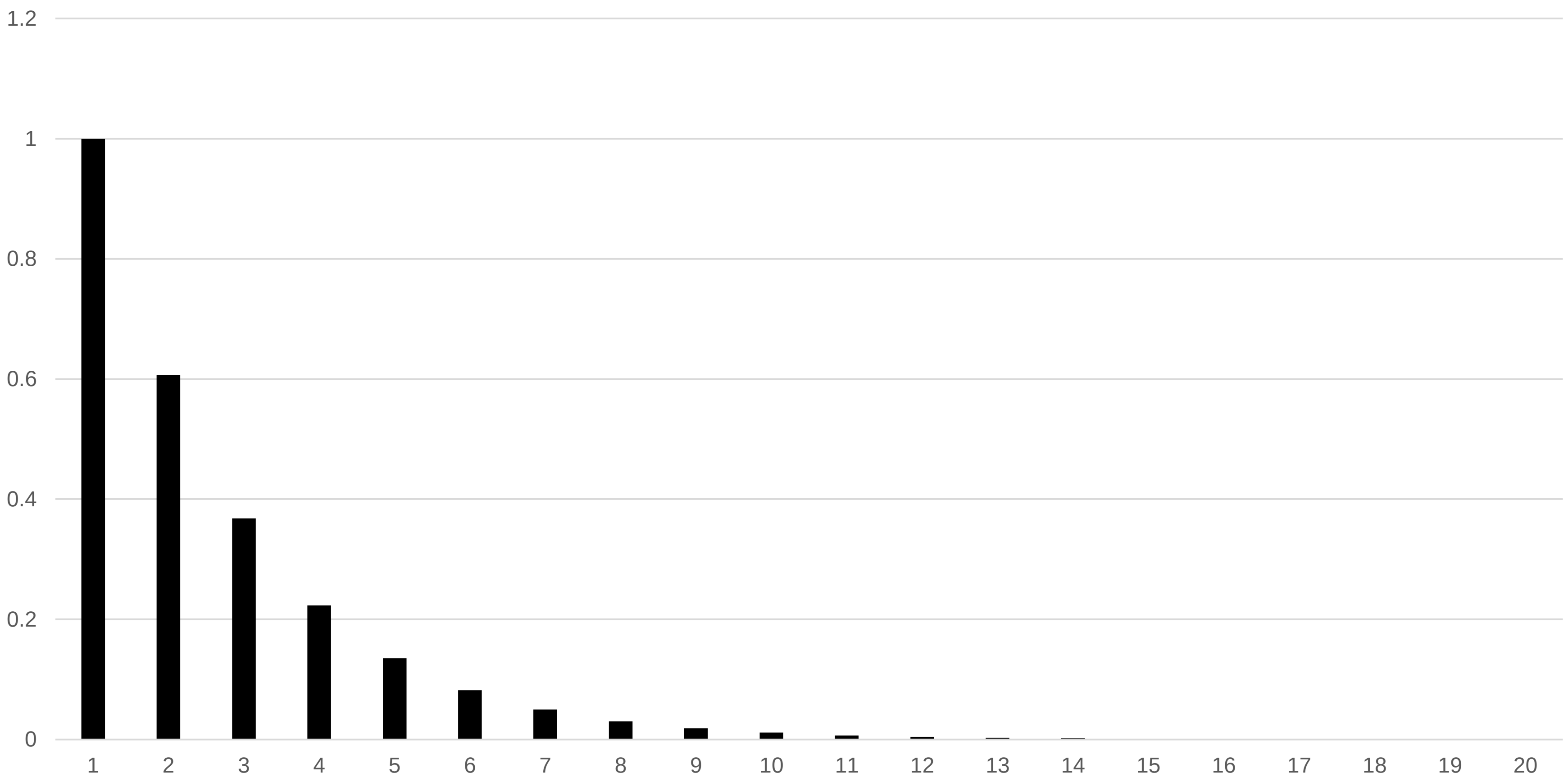
Bottom Left: The actual normal distribution (with die distributions superimposed.)

https://en.wikipedia.org/wiki/File:Dice_sum_central_limit_theorem.svg

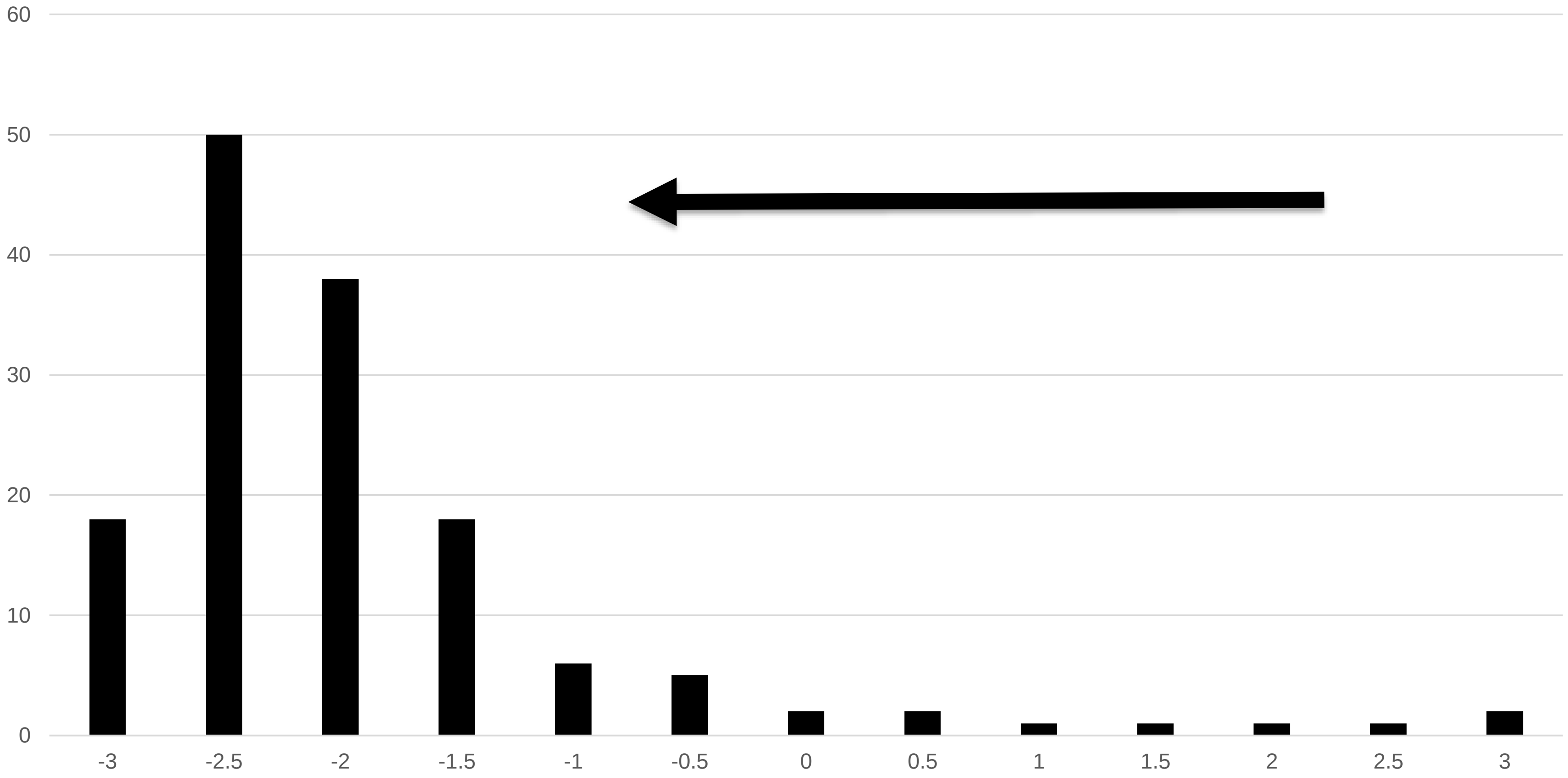
Uniform Distribution



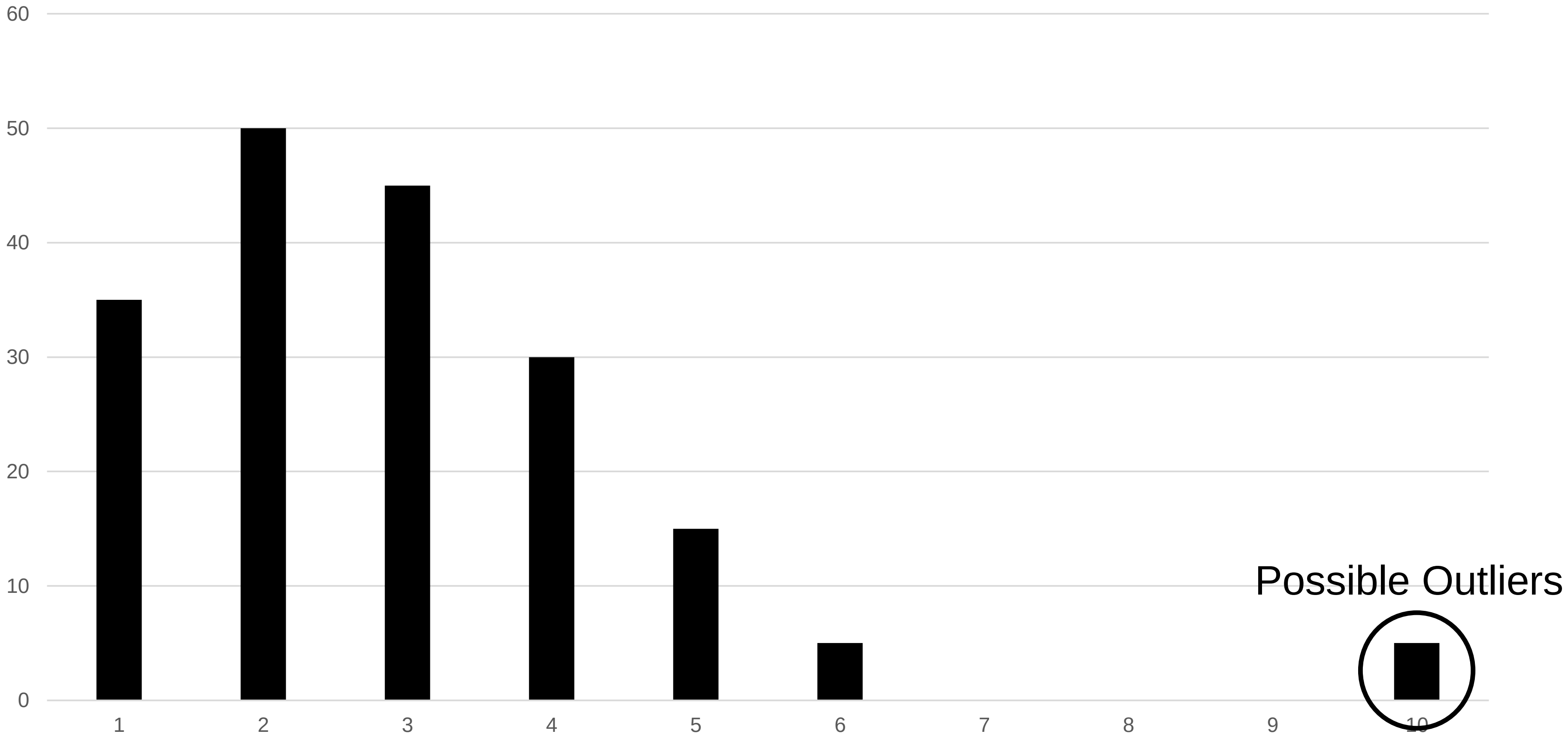
Exponential Distribution



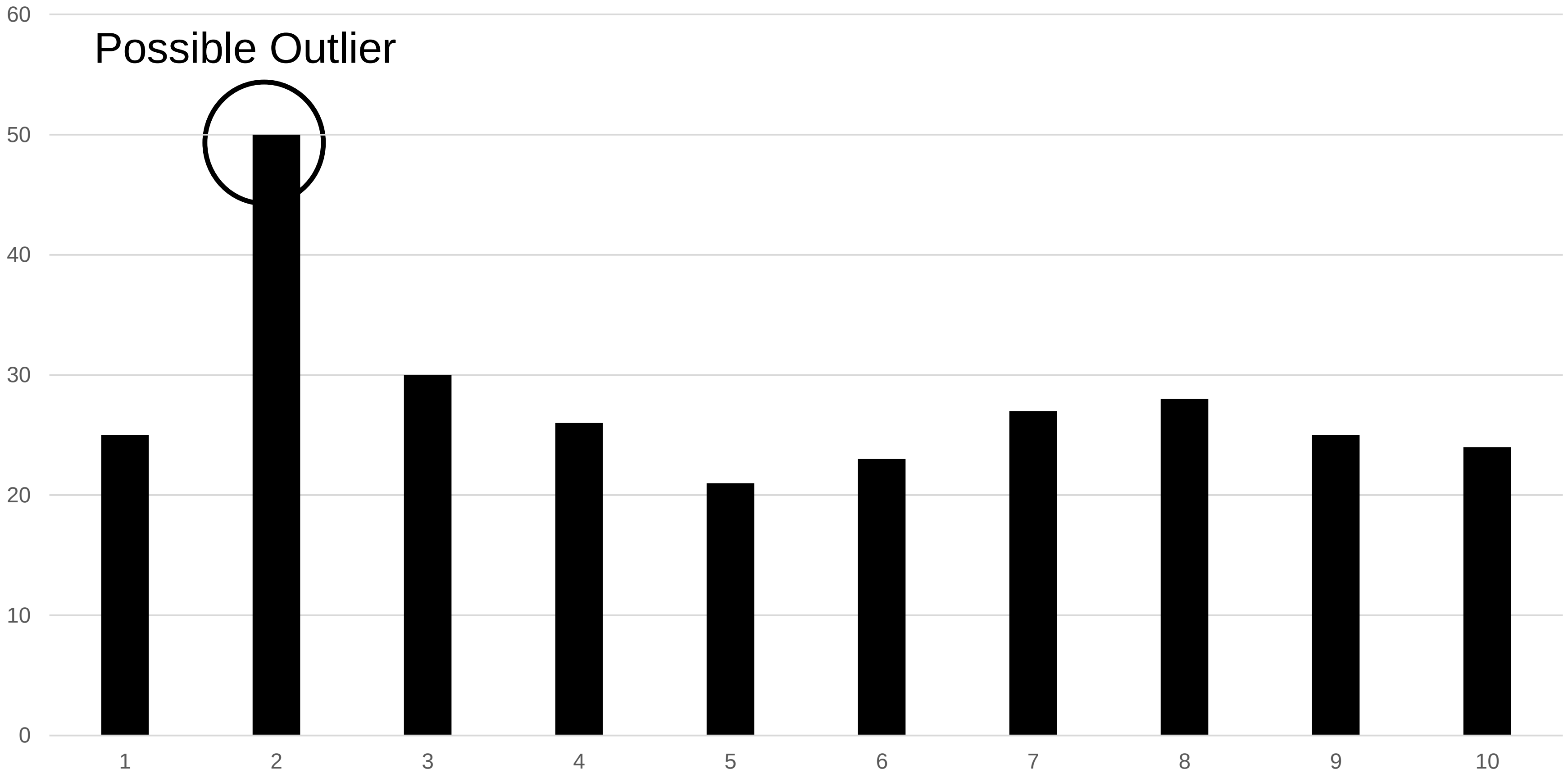
A Skewed distribution



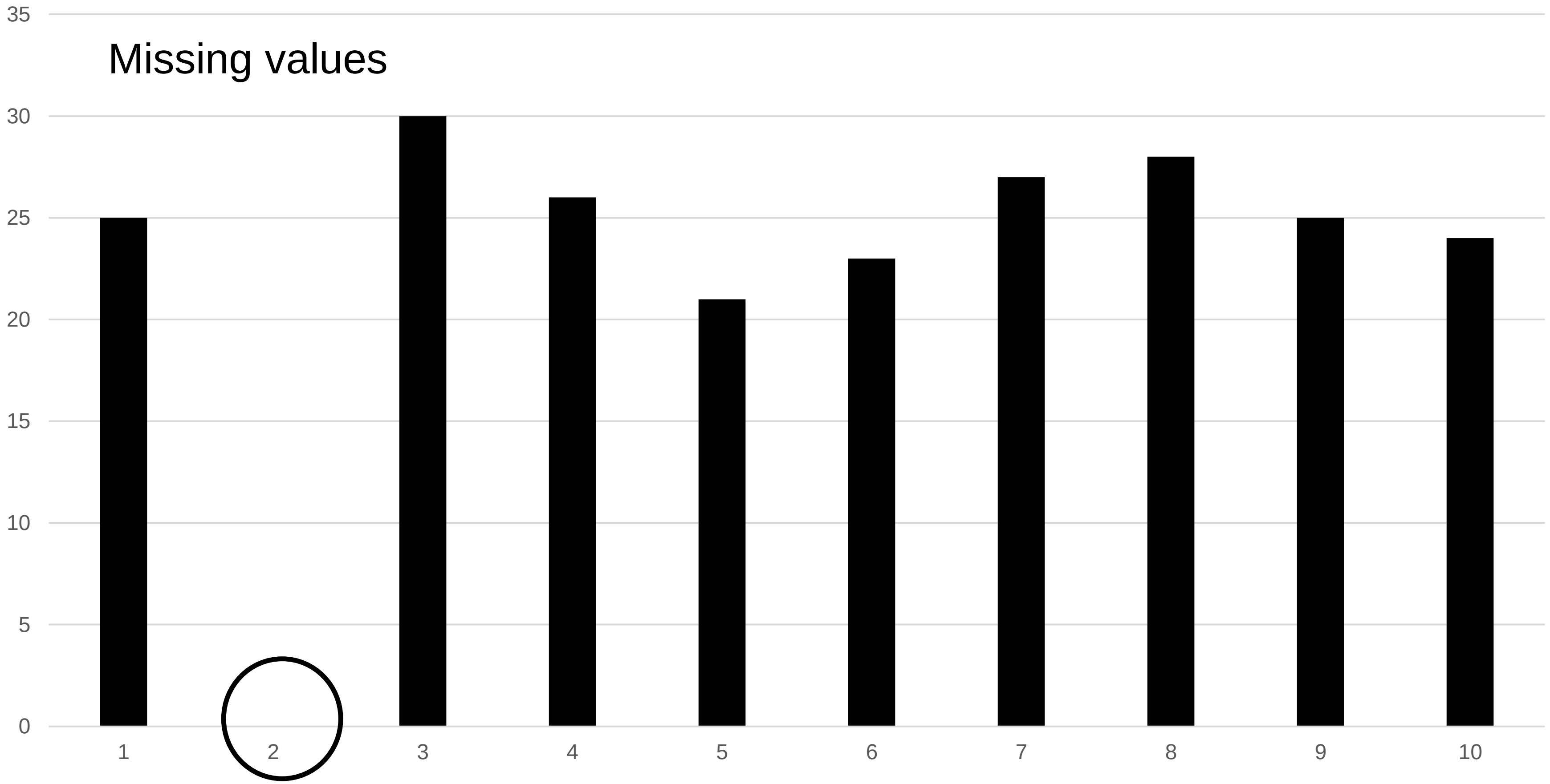
Outliers



Outliers



Outliers



Histograms: Number of bins

Number of bins according to Sturges' rule:

$$k = \lceil \log_2(n) + 1 \rceil$$

where n is the sample size.

(Sturges' rule is suitable for data from normal distributions and from data sets of moderate size.)

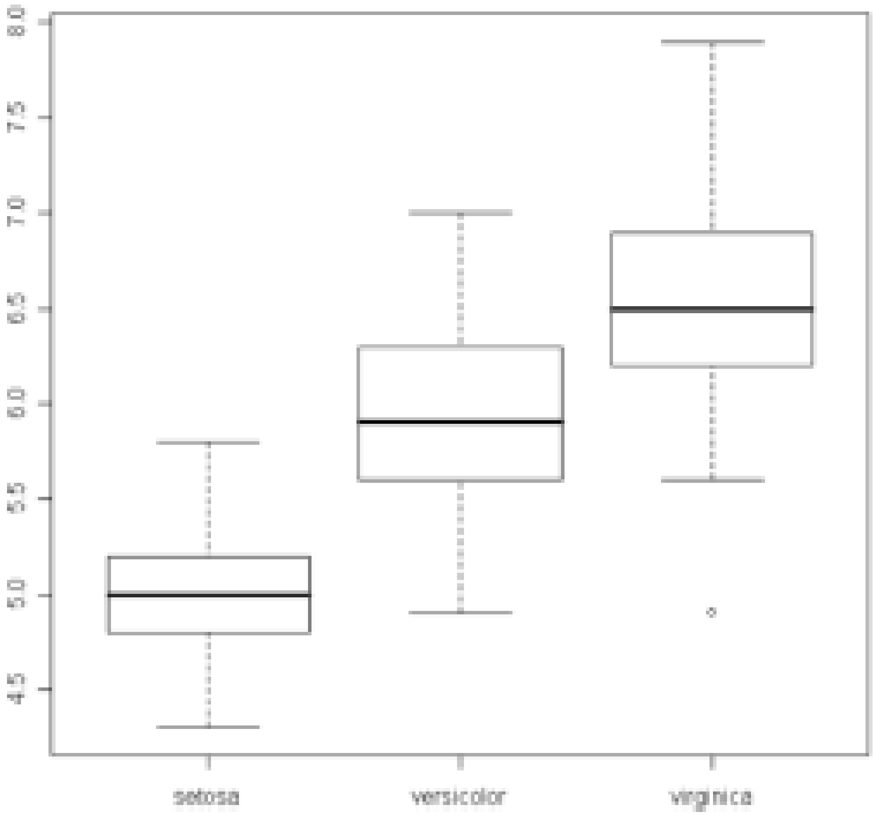
Box plots

- Histograms will show the overall shape of a variable's distribution
  - A coarse overview.
  - Major problems will appear, individual outliers hard to spot.

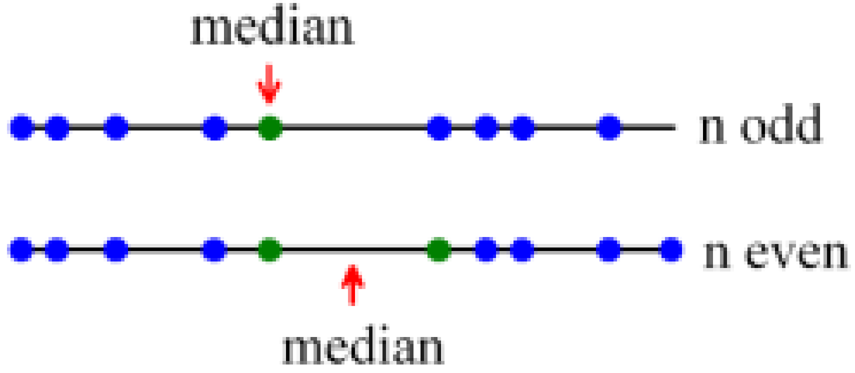
- Box plots show the centre and range (standard deviation or standard error of data)
  - Individual outliers are clearly shown as dots or stars.

Boxplots

- Each boxplot shows the *distribution* of a single attribute's values
- Each boxplot represents a species; the boxplots themselves show the distributions of another attribute value, allowing us to compare the distributions for different species.



Reminder: Median, quantiles, quartiles, interquartile range



Median: The value in the middle (for the values given in increasing order).

q%-quantile $(0 < q < 100)$: The value for which q% of the values are smaller and 100-q% are larger.
The median is the 50%-quantile.

Quartiles: 25%-quantile (1st quartile), median (2nd quantile), 75%-quantile (3rd quartile).

Interquartile range (IQR): 3rd quantile - 1st quantile.

Even more basic reminder: Mean, Median and Mode

Given the following 17 numbers, find the mean, median and mode:
4.3, 5.1, 3.9, 4.5, 4.4, 4.9, 5.0, 4.7, 4.1, 4.6, 4.4, 4.3, 4.8, 4.4, 4.2, 4.5, 4.4

Mean: the average of the numbers ie. 76.5/17 = **4.5**

Median: to calculate the median, we must first put the set in ascending order:
3.9, 4.1, 4.2, 4.3, 4.3, 4.4, 4.4, 4.4, 4.4, 4.5, 4.5, 4.6, 4.7, 4.8, 4.9, 5.0, 5.1

We want to find the middle value: in this case the middle value is the ninth value:
3.9, 4.1, 4.2, 4.3, 4.3, 4.4, 4.4, 4.4, **4.4**, 4.5, 4.5, 4.6, 4.7, 4.8, 4.9, 5.0, 5.1
(NB if there is an even number of values, you need to find the mean of the middle two values)

Mode: the most frequently occurring value – in this case it is also **4.4**.

Even more basic reminder: quartiles

We divide our dataset at three points: the median, and the middle points of the two halves: this divides the entire dataset into quarters – “quartiles”

The top point of the first quarter is Q_1, the median value is Q_2, the middle value for the second half of the set is Q_3, and Q_4 is the largest value.

1. Find the median value: $\mathbf{Q_2}$ = **4.4**

3.9, 4.1, 4.2, 4.3, 4.3, 4.4, 4.4, 4.4, **4.4**, 4.5, 4.5, 4.6, 4.7, 4.8, 4.9, 5.0, 5.1

2. We remove the median value and now have two sets of eight values each:

3.9, 4.1, 4.2, 4.3, 4.3, 4.4, 4.4, 4.4 and 4.5, 4.5, 4.6, 4.7, 4.8, 4.9, 5.0, 5.1

We find the median value for each of these sets:

3.9, 4.1, 4.2, 4.3, 4.3, 4.4, 4.4, 4.4 $\mathbf{Q_1}$ = (4.3 + 4.3)/2 = **4.3**

4.5, 4.5, 4.6, 4.7, 4.8, 4.9, 5.0, 5.1 $\mathbf{Q_3}$ = (4.7 + 4.8)/2 = **4.75**

3. We find the largest value in the list: $\mathbf{Q_4}$ **= 5.1**

Source: http://www.purplemath.com/modules/boxwhisk.htm

Reminder: quartiles, continued:

- **first quartile** (Q_1) = **lower quartile** = 25th percentile
  (splits off the lowest 25% of data from the highest 75%)

- **second quartile** (Q_2) = median = 50th percentile
  (cuts data set in half)

- **third quartile** (Q_3) = **upper quartile** = 75th percentile
  (splits off the highest 25% of data from the lowest 75%)

- The difference between the upper and lower quartiles is called the *interquartile range*.

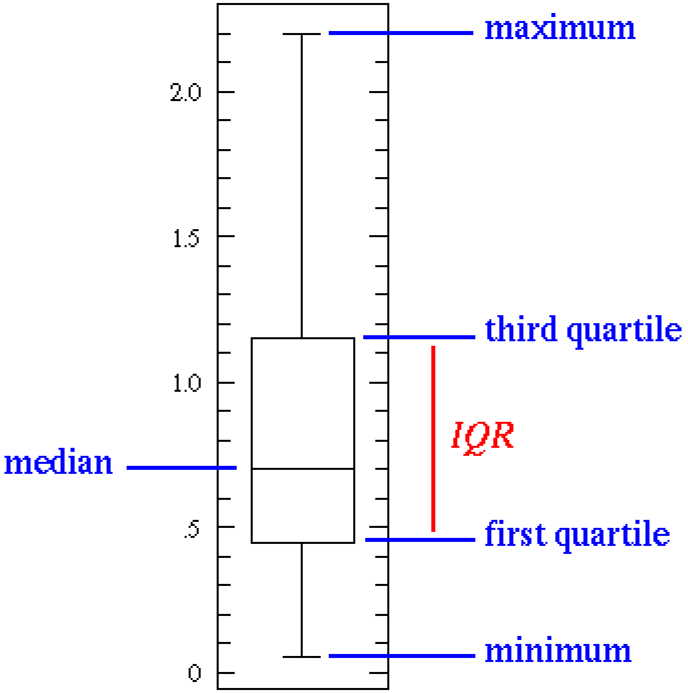
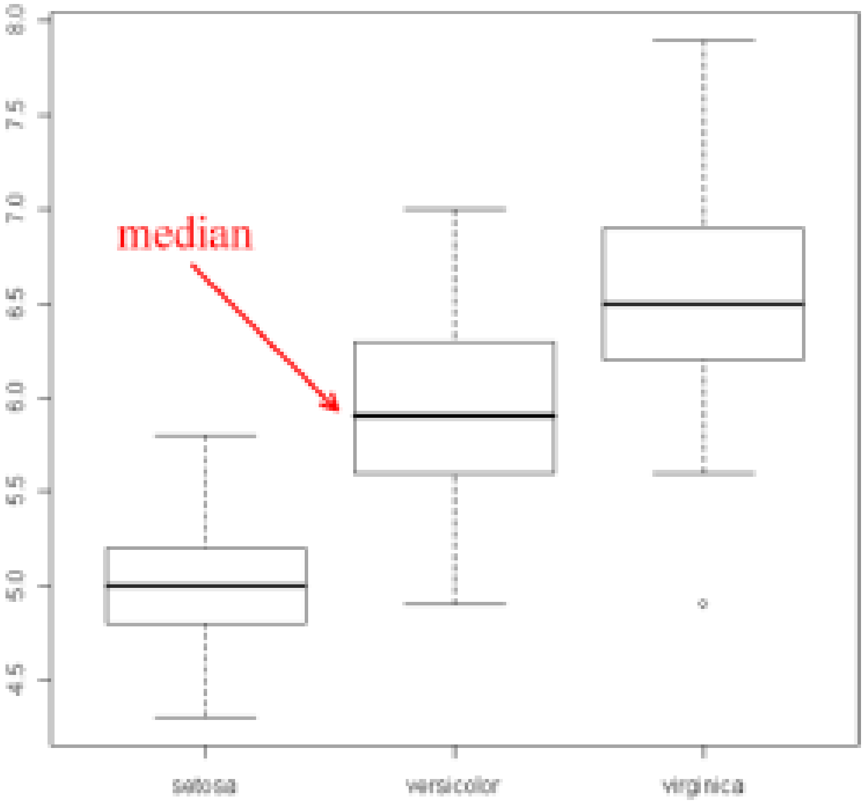


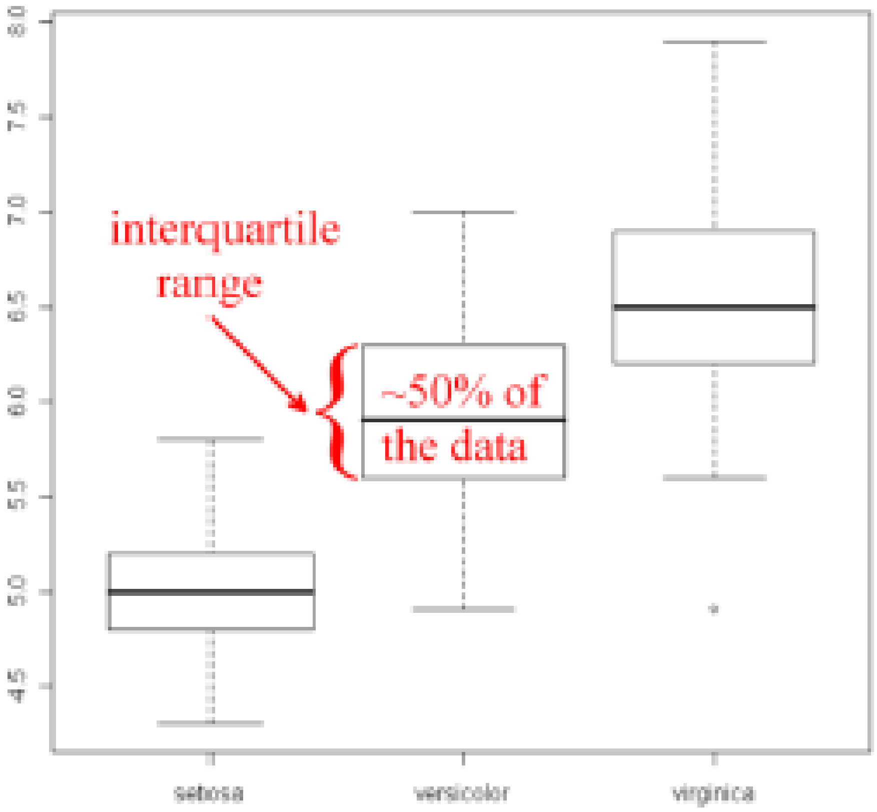
diagram: http://www.physics.csbsju.edu/stats/box2.html

Source: http://en.wikipedia.org/wiki/Quartile

Boxplots

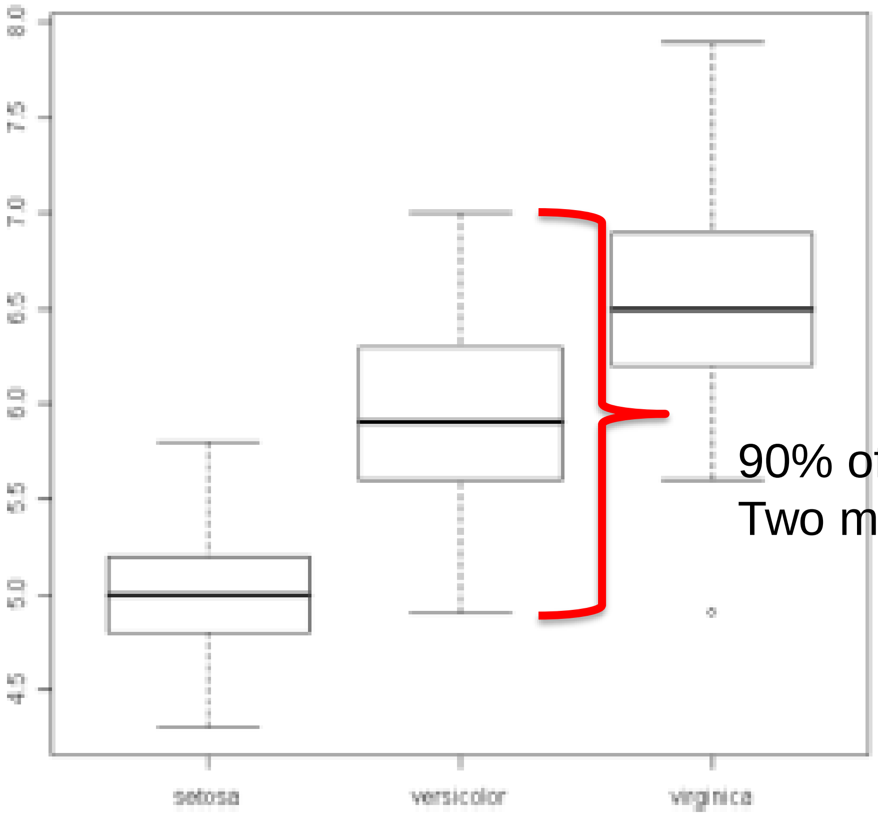


Boxplots


8.0
7.5
7.0
6.5
6.0
5.5
5.0
4.5
setosa
versicolor
virginica
interquartile range
~50% of the data

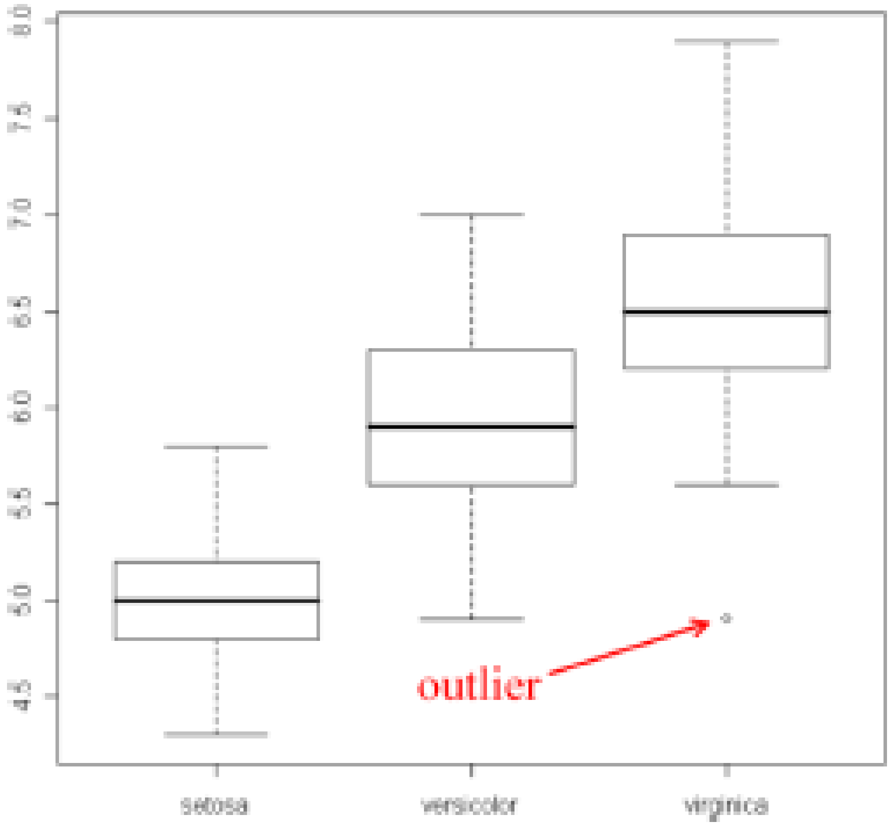
Boxplots

- Each boxplot shows the *distribution* of a single attribute's values
- Each boxplot represents a species; the boxplots themselves show the distributions of another attribute value, allowing us to compare the distributions for different species.

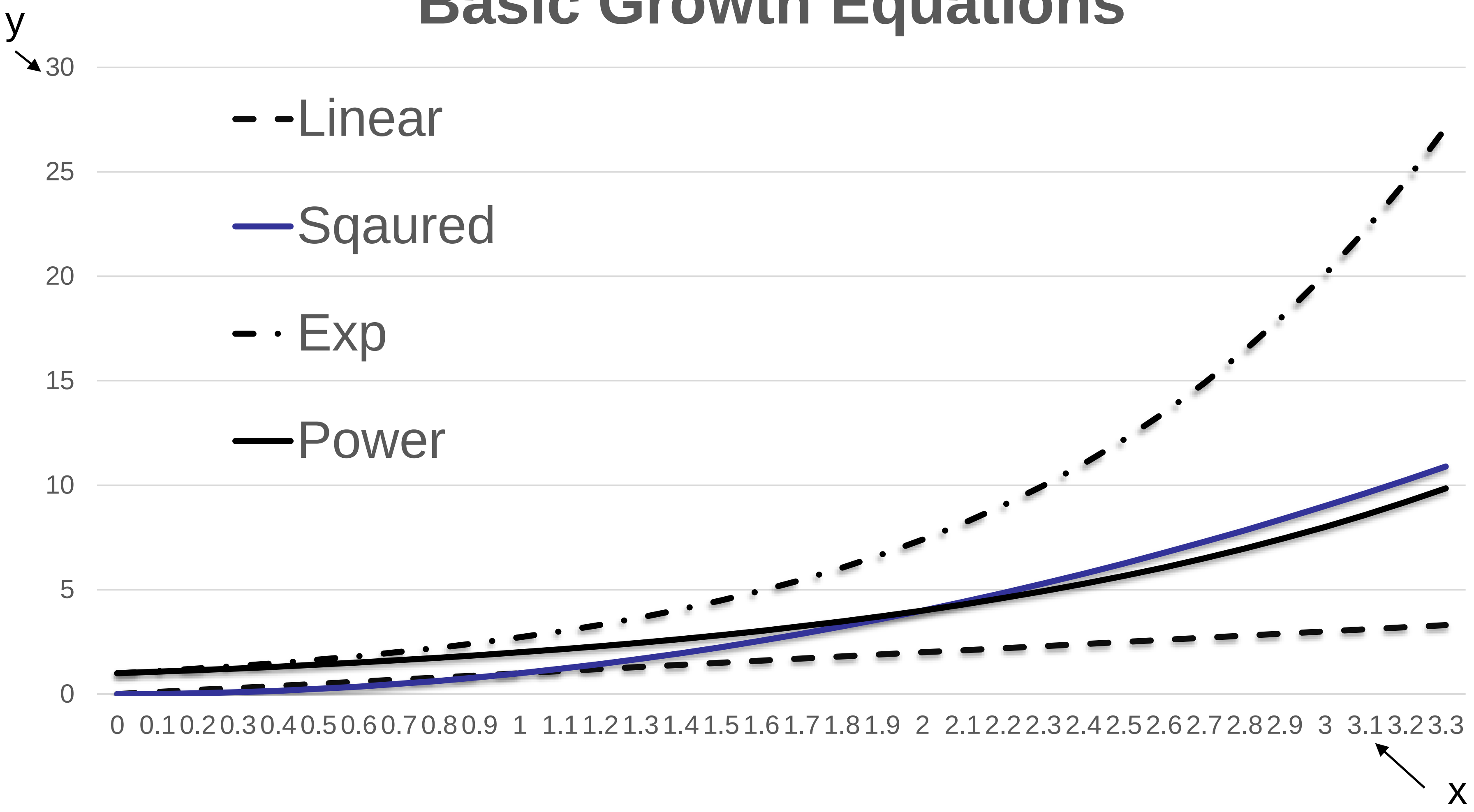


90% of data between these
Two marks.

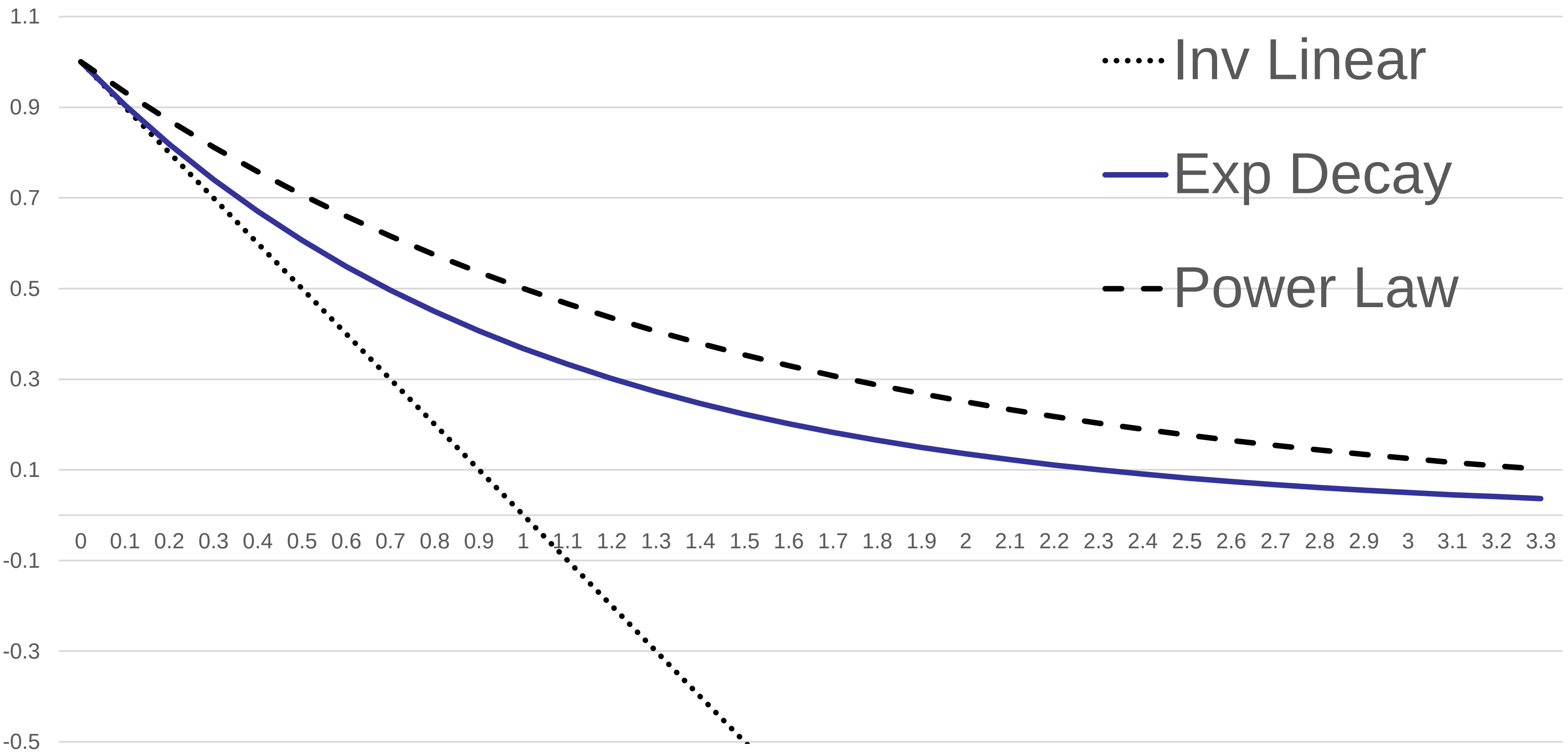
Boxplots



Basic Growth Equations



Decay Equations



Many names

Name	Name	Name	Type	Shown
Linear	Correlated	Linear Regression	Growth	First Slide
Linear	Correlated	Linear Regression	Decay	Second Slide *
Exponential Decay	Exponential distribution	Laplacian distribution	Decay	Second Slide
Power Law	Inverse Power equation		Decay	Second Slide
Increases Exponentially	Exponential equation		Growth	First Slide

Choosing a Visualisation

How many variables?

1. Histogram, Frequency bar chart, Box plot
2. Scatterplot, Bar chart
3. Scatterplot + Colour/shape, Parallel co-ordinates, Box-plot

Many. Matrix, Parallel co-ordinates, Box-Plot

One ‘dimension’ per Variable

Choosing a Visualisation

What type of Variables

Categorical: Frequency bar chart

Interval/Ratio: Histogram

Category & Interval/Ratio: Bar Chart

Interval/Ratio & Interval/Ratio: Scatterplot

Category & Interval/Ratio & Interval/Ratio : Coloured Scatterplot

Interval/Ratio & Interval/Ratio & Interval/Ratio etc. Parallel coordinates

Choosing a Visualisation

What type of Variables

Generally

Categorical -> Colour or Shape

Interval/Ratio -> Position in space (x or y co-ordinates)

Finding Patterns/Problems

Finding outliers:

Single variable: Histograms, Frequency bar charts, box plots

2 or more Variables: Scatterplots, Box Plot

Many: Multi-dimensional scaling or PCA (see Berthold 2010)

Finding clusters

2 or more Variables: Scatterplots

Many: Multi-dimensional scaling or PCA (see Berthold 2010)

Finding relationships

2 variables: Scatterplots

Many: Matrix of Scatterplots

VISUALISATION DOES NOT REPLACE STATISTICAL TESTS

A checklist for data understanding

- Determine the quality of the data. (e.g. syntactic accuracy)
- Find outliers. (e.g. using visualization techniques)
- Detect and examine missing values. Possible hidden by default values.
- Discover new or confirm expected dependencies or correlations between attributes.
- Check specific application dependent assumptions (e.g. the attribute follows a normal distribution)
- Compare statistics with the expected behavior.

Recommended Reading

“Guide to Intelligent Data Analysis: How to Intelligently Make Sense of Real Data”, M Berthold, Michael, Christian Borgelt; Frank Höppner; F Klawonn

Revise chapter 2. pp 15-23

Read chapter 4 pp 333-80

Some useful resources

Visualisation repositories:

- Tableau's visual gallery http://www.tableausoftware.com/learn/gallery
- D3 gallery: https://github.com/mbostock/d3/wiki/Gallery

Data specific

- Temporal data: http://timeviz.net/
- Graph data: http://www.visualcomplexity.com/vc/
- Trees: http://www.treevis.net

References

Tukey, J.W. (1977) Exploratory Data Analysis. Addison-Wesley, Reading

Compendium slides for Guide to Intelligent Data Analysis, Springer 2011. Michael R. Berthold, Christian Borgelt, Frank H▸ oppner, Frank Klawonn and Iris Ad. http://www.informatik.uni-konstanz.de/gidabook/teaching-material/?print=1