Data analysis methods in weather and climate research

Dr. David B. Stephenson
Department of Meteorology
University of Reading
www.met.rdg.ac.uk/cag

2. Exploratory Data Analysis (EDA)

- Data tabulation
- Summary plots
- Measures of location, scale, and shape
- Rank statistics and empirical quantiles
- Transformation of data values

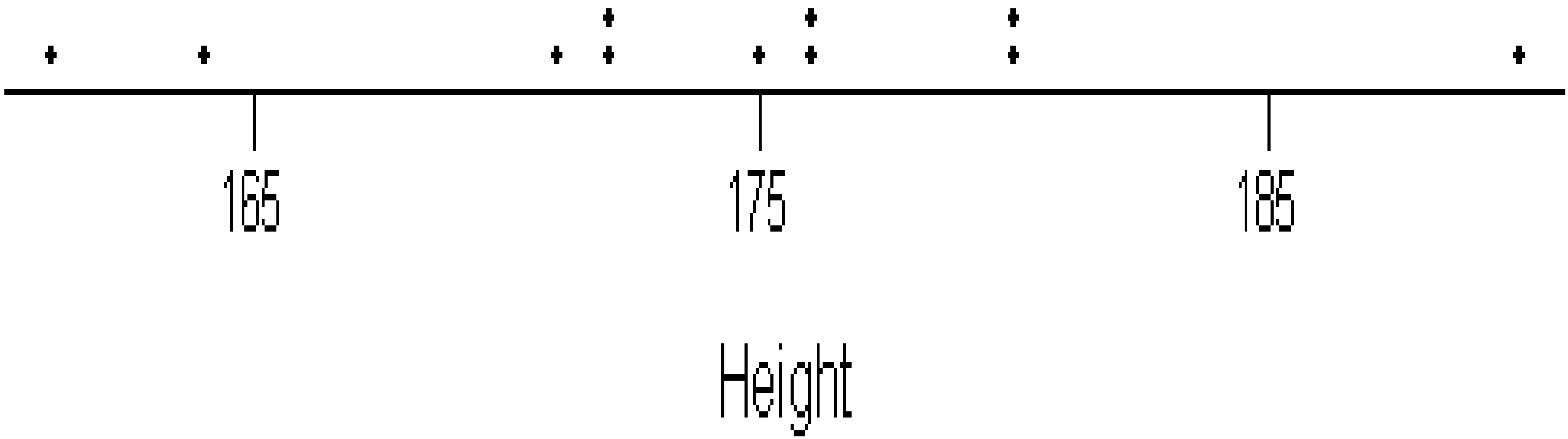
2. Data tabulation

Object	Age (years)	Height (cm)	Weight (kgs)
1	30.9	180	76
2	26.9	164	64
3	33.2	176	87
4	28.5	172	75
5	32.3	176	75
6	37.0	180	86
7	38.3	171	65
8	31.5	172	76
9	32.8	161	75
10	37.7	175	85
11	29.1	190	83

rdgmorph.txt
Meteorologist dat

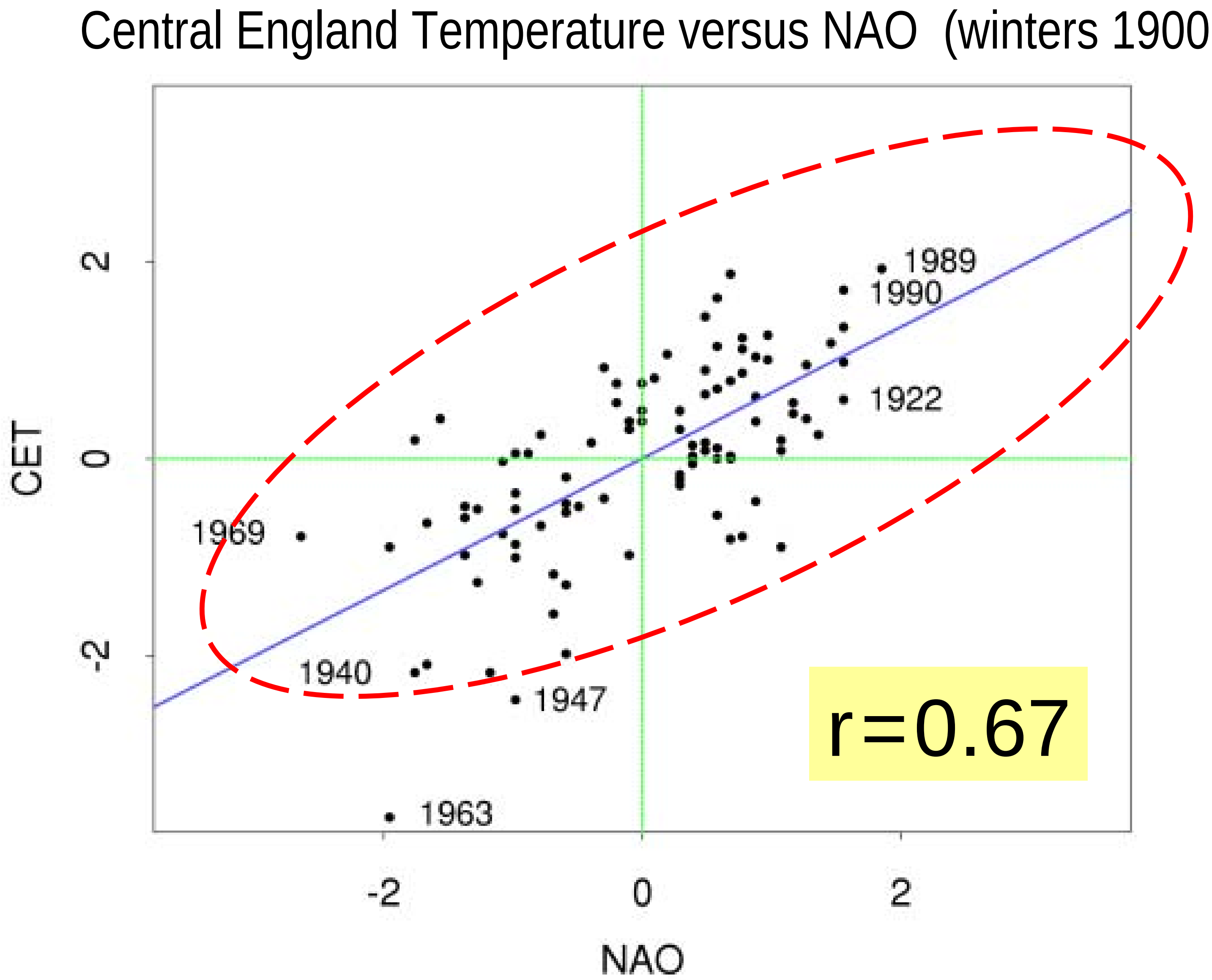
Rows=objects
Columns=variabl
Sample size=n=1

2. Dotplot (=1-d scatter plot)

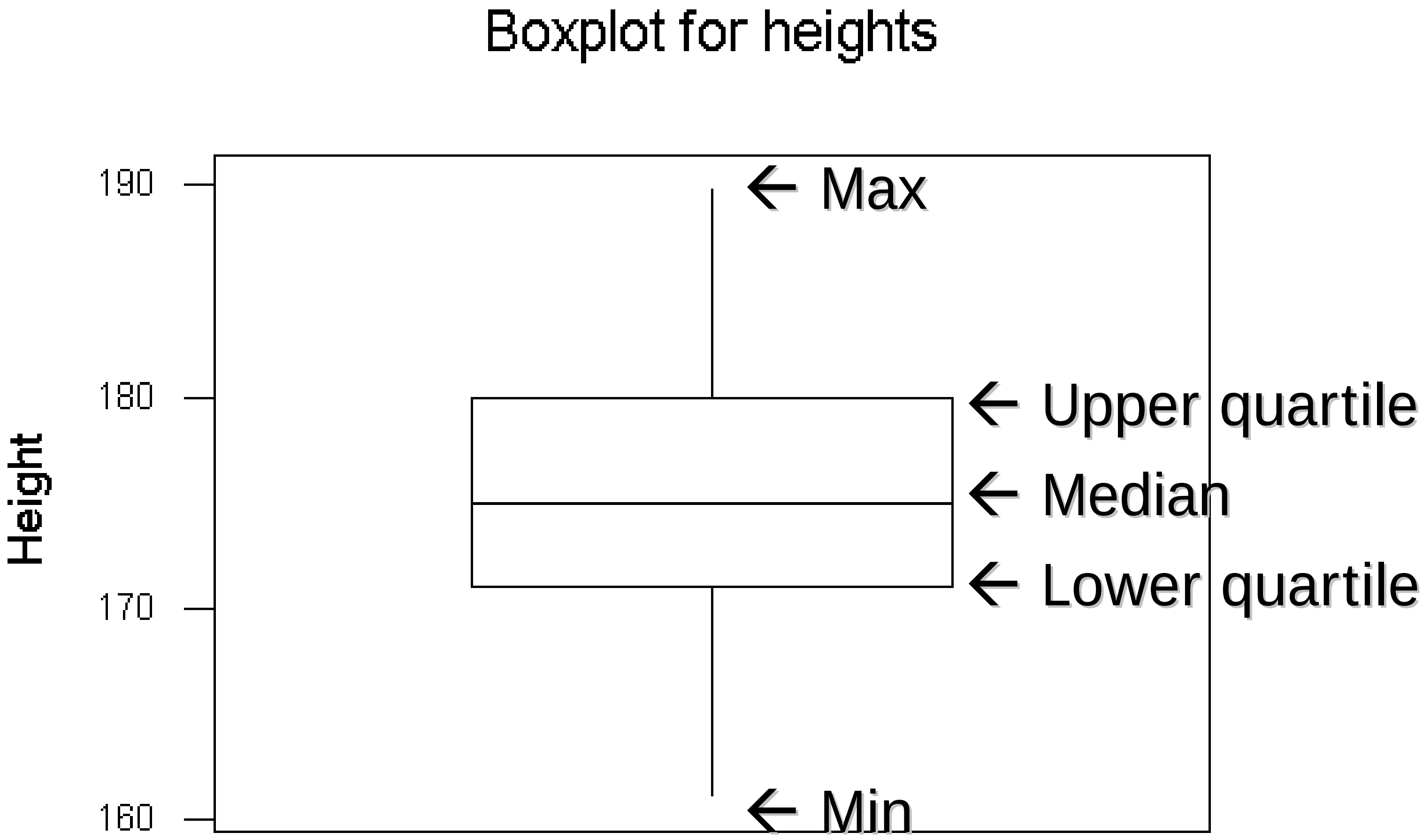


Note the “tied” values that occur in this small sample

2. Scatter plot (2-d dotplot)

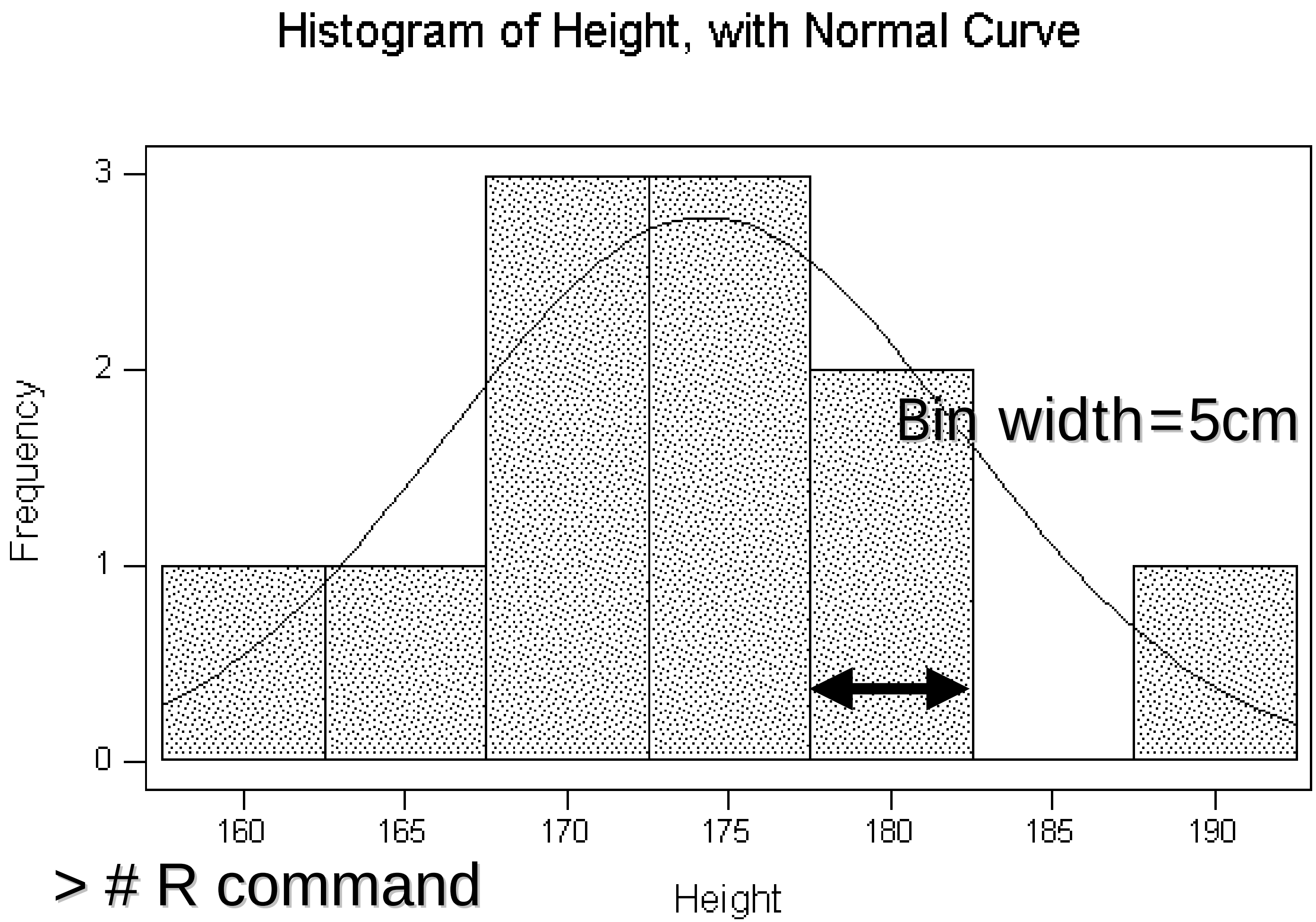


2. Boxplot




```
> # R command
> boxplot(x)
```

2. Histogram

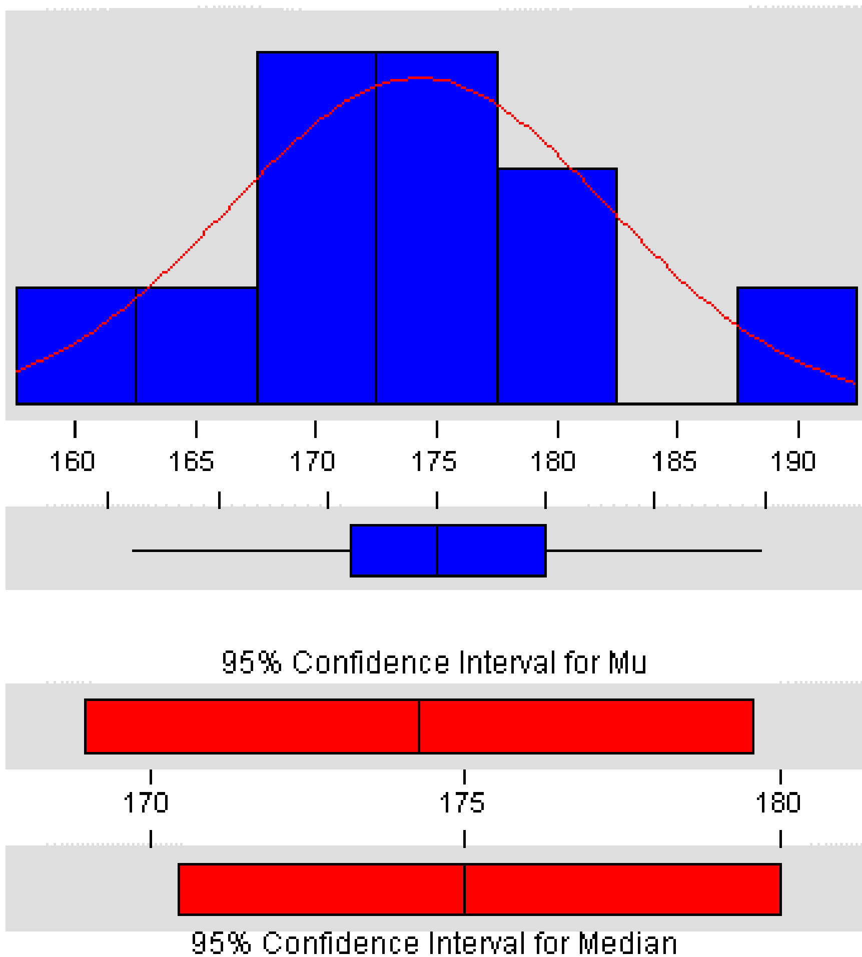


> # R command

> hist(x)

2. Descriptive summary

Descriptive Statistics



Variable: Height

Anderson-Darling Normality Test

A-Squared:	0.291
P-Value:	0.541
Mean	174.273
StDev	7.888
Variance	62.2182
Skewness	0.198168
Kurtosis	0.838718
N	11
Minimum	161.000
1st Quartile	171.000
Median	175.000
3rd Quartile	180.000
Maximum	190.000

95% Confidence Interval for Mu

168.974 179.572

95% Confidence Interval for Sigma

5.511 13.843

95% Confidence Interval for Median

170.425 180.000

2. Summary measures

- Centre/Location
  - Mean (m or x)
  - Median (x)
- Scale/Spread
  - Standard deviation (s)
  - Interquartile range (IQR)
- Shape
  - Skewness (e.g. b1)
  - Kurtosis (e.g. b2)

2.3 The sample mean

$$\bar{x} = \frac{x_1 + x_2 + \ldots + x_n}{n} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

Standard measure of the central location of the data.

```
> # R command
> mean(x)
```

2.3 The standard deviation

$$s = \sqrt{\overline{(x-\overline{x})^2}} = \sqrt{\overline{x^2} - (\overline{x})^2}$$

Std. Deviation=root mean squared deviation
Standard measure of the spread/scale of the data.

```
> # R command
> sd(x)
```

2.3 Higher moments about mean

$$m_r = \overline{(x - \overline{x})^r}$$

Give information about the shape of the distribution
e.g. all odd moments are zero for a symmetric distribution

```
> # R command to do m4
> mean((x-mean(x))^4)
```

2.3 Skewness and kurtosis

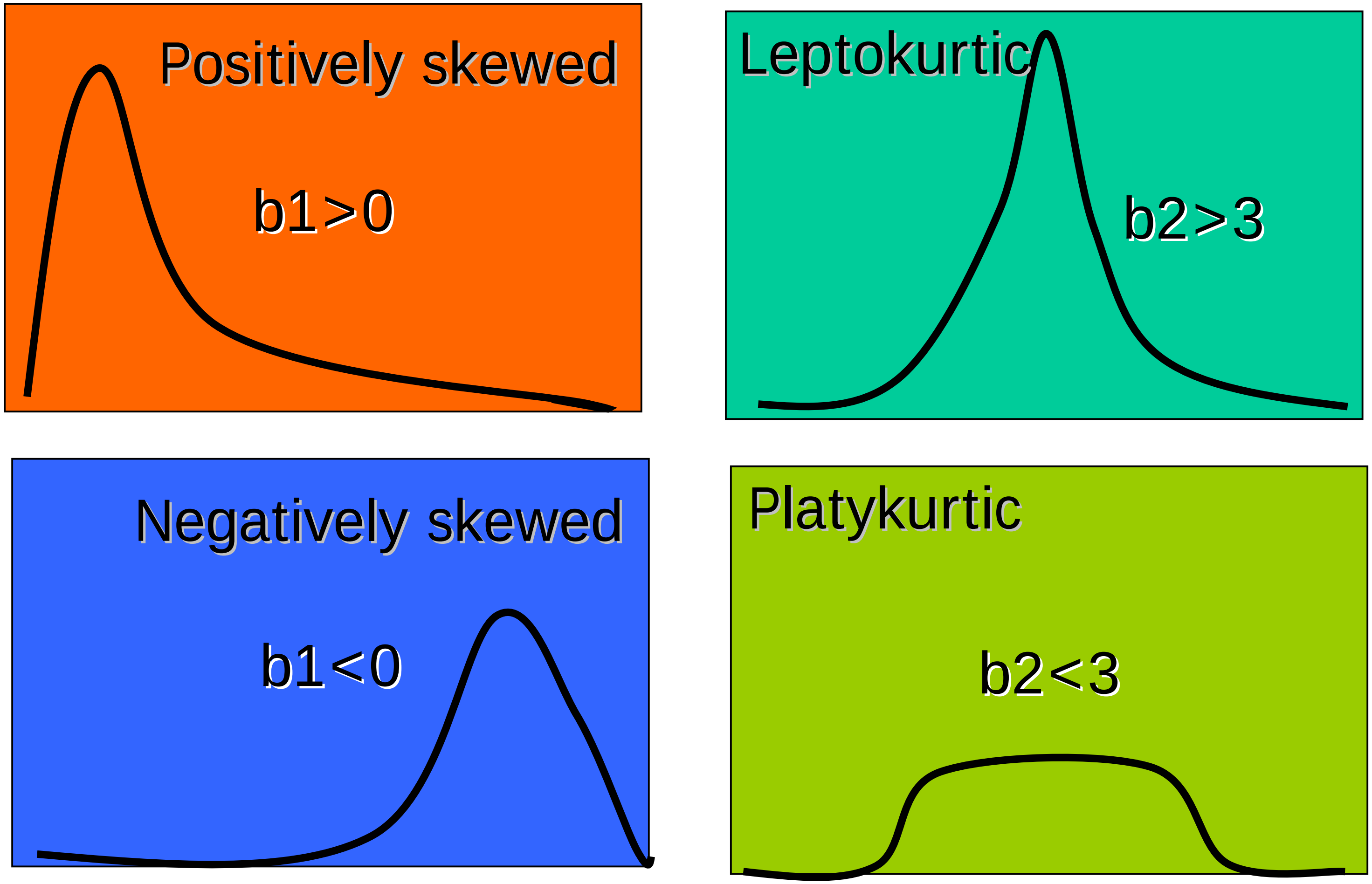
$$Skewness = b_1 = m_3 / s^3$$

$$Kurtosis = b_2 = m_4 / s^4$$

For normal (Gaussian) distribution:
Skewness=0 (symmetric) Kurtosis=3
Kurtosis > 3 “leptokurtic” (fat tails and sharp peak)
Kurtosis < 3 “platykurtic” (thin tails and flatter peak)

```
> # R commands
> b1=mean(scale(x)^3)
> b2=mean(scale(x)^4)
```

2.3 Shapes of distributions



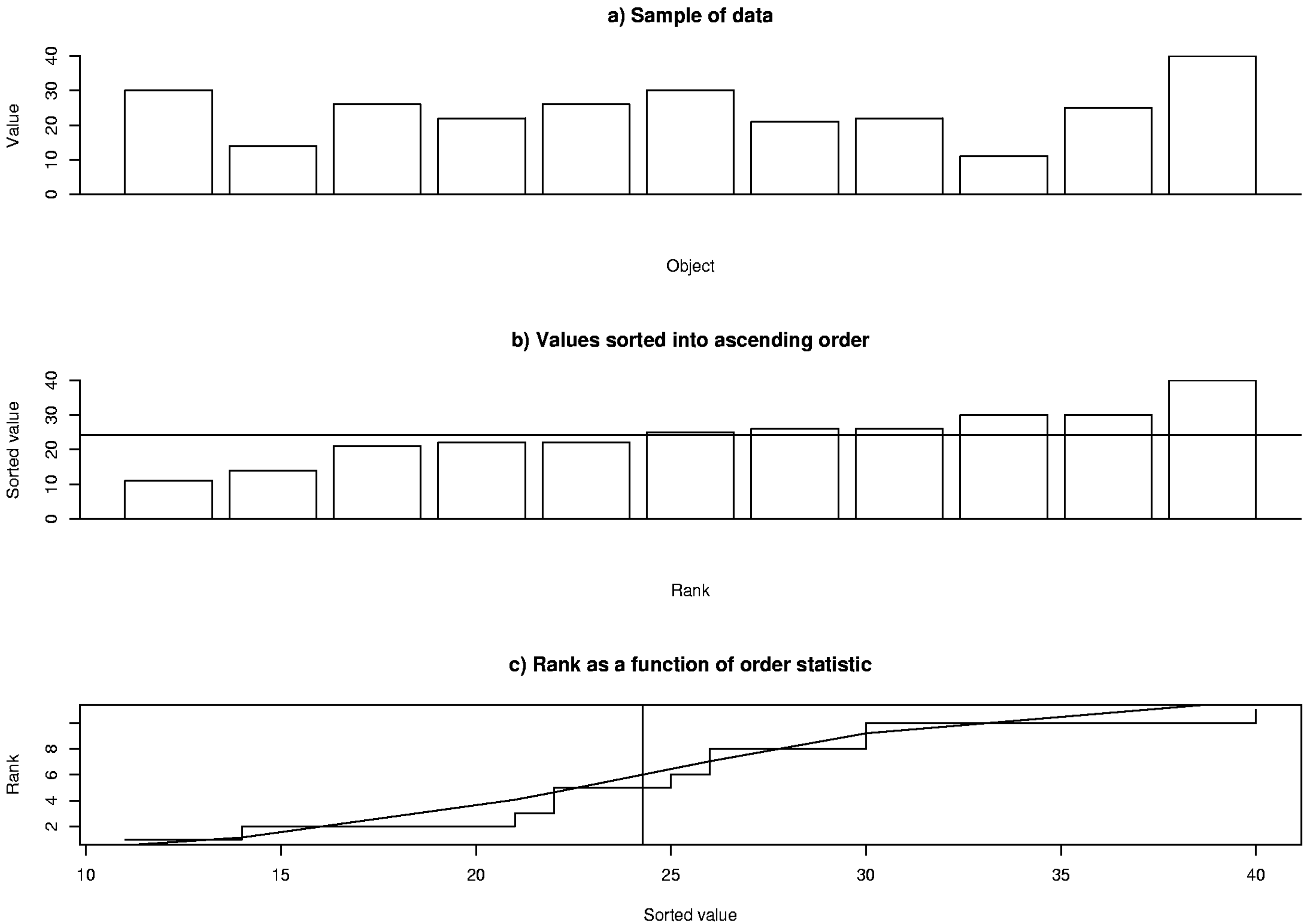
Resistant and Robust statistics

- **Resistant statistic** – one that is not overly sensitive to small or large outlier data

  e.g. IQR compared to max-min range.

- **Robust statistic** – one that is not dependent on the details of the probability distribution e.g. rank-statistics (median etc.)

2.4 Empirical cumulative distribution



2.4 Ranks

Object	Height (cm)	Rank r
1	180	9= → UPPER QUARTILE
2	164	2
3	176	7=
4	172	4=
5	176	7=
6	180	9=
7	171	3 → LOWER QUARTILE
8	172	4=
9	161	1 → MIN
10	175	6 → MEDIAN
11	190	11 → MAX

rdgmorph.txt
Meteorologist dat

Rows=objects
Columns=variable
Sample size=n=1

2. Empirical quantiles

rank r

order statistic $x_{[r]}$

empirical probability $p = \dfrac{r}{n+1}$

empirical p'th quantile :

$$x_p = \begin{cases} x_{[(n+1)p]} & \text{if (n+1)p integer} \\ 0.5*(x_{[(n+1)p]} + x_{[(n+1)p+1]}) & \end{cases}$$

```
> # R command
> rank(x)
> sort(x)
> order(x)
> quantile(x,0.9)
```

2.5 Transformation of data

- "Center"

  remove sample mean

$$y = x - \bar{x}$$

- "Standardize"

  remove sample mean and scale

$$y = \frac{x - \bar{x}}{s_x}$$

- "Normalize"

  nonlinear transformation

$$y = \frac{x^\nu - 1}{\nu}$$

$$0 \le \nu \le 1$$

Summary

- EDA
  - *know exactly how the data was produced*
  - *try to get hold of the raw untreated data*
  - *let the data speak for themselves*
- Summarise location, scale, shape
- Investigate outliers, tied values, and any other strange features
- Transform the data if necessary