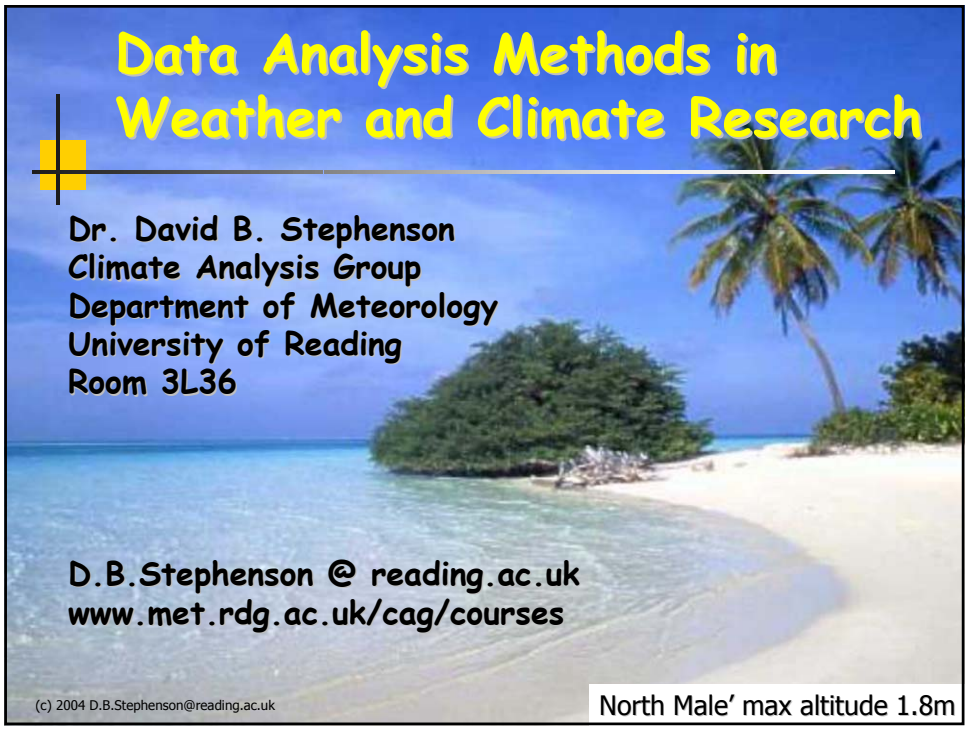
Data Analysis Methods in Weather and Climate Research

Dr. David B. Stephenson
Climate Analysis Group
Department of Meteorology
University of Reading
Room 3L36

D.B.Stephenson @ reading.ac.uk
www.met.rdg.ac.uk/cag/courses



North Male' max altitude 1.8m

This course will help to develop

- Your appreciation of what statistics is really about
- Your skills at analyzing data using statistical software
- Your ability to choose and apply appropriate methods
- Your ability to correctly interpret the analysis results
- Your capability to learn more about statistics …

Statistics: a great cure for insomnia

This course will not …

- Teach you EVERYTHING about statistics !
- Give you cookbook recipes for success
- Show you how to perform miracles with data
- Be the last thing you learn in statistics …

Prerequisites

- Some knowledge of mathematics
- Some experience of spreadsheet software (e.g. excel)
- + lots of curiosity !

Activities

- Lectures 9-11am weeks 7-10
- Practical classes 2-5pm weeks 7-10
- Self-study of notes/books/web

Further reading …

1. Hand out notes – please spend a couple of hours each week going through each chapter. Also web based material: www.met.rdg.ac.uk/courses/stats
2. M.R. Spiegel (1992) "Theory and problems of statistics", Schaum ouline series.
3. M.H. DeGroot and M.J. Schervish (2002) Probability and Statistics, (3rd edition), Addison-Wesley, 816 pp.
4. D. Wilks (1995) "Statistical Methods in the Atmospheric Sciences", Academic Press
5. H. von Storch and F. Zwiers (1999) "Statistical analysis in climate research", Cambridge University Press.

Assessment

100% on written up practical data analysis assignment that covers all the main parts of the course. To be handed in by Friday 26 March 2004.

Course outline

1. Introduction
2. Descriptive sample statistics
3. Basic probability concepts
4. Probability distributions
5. Parameter estimation
6. Statistical hypothesis testing
7. Basic linear regression
8. More advanced regression
9. Introduction to time series

1. Introduction

1. Brief history of statistics and probability
2. Why "statistics" is more than just "data analysis"
3. Descriptive versus inferential statistics
4. Statistical software

Definition of statistics

Statistics – first applied to the political science concerned with the facts of a state or a community XVIII; all derived immediately from German *statistisch* adj., *statistik sb.*; whence *statistician XIX.*

History of Statistics and Probability

1500-1599	Sir W. Petty	
1600-1699	J. Graunt E. Halley	B. Pascal P. de Fermat
1700-1799		J. Bernoulli C. Huygens A. DeMoivre Rev. Bayes
1800-1899	C.F. Gauss F. Galton	P.S. Laplace
1900-1999	K. Pearson Student aka W. Gosset R. Fisher + many others	Kolmogorov + many others

A generall Bill for this present year, ending the 19 of *December* 1665. according to the Report made to the KINGS most Excellent Majesty.

By the Company of Parish Clerks of *London*, &c.

Buried in the 97 Parishes within the walls, — 15207 Whereof, of the Plague — 9887

Buried in the 16 Parishes without the walls — 41351 Whereof, of the Plague — 28888

Buried in the 12 out-Parishes, in Middlesex and Surrey — 18554 Whereof, of the Plague — 21420

The Total of all the Christnings — 9967

The Total of all the Burials this year — 97306

Whereof, of the Plague — 68596

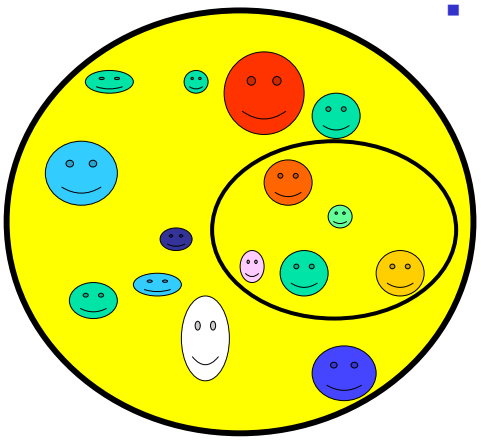
The Diseases and Casualties this year.

Abortive and Stilborne	617	Executed	21	Palsie	30
Aged	1545	Flox and Small Pox	655	Plague	68596
Ague and Feaver	5257	Found dead in streets, fields, &c.	20	Plannet	6
Appoplex and Suddenly	116	French Pox	86	Plurisie	15
Bedrid	10	Frighted	23	Poysoned	1
Blasted	5	Gout and Sciatica	27	Quinsie	35
Bleeding	16	Grief	46	Rickets	557
Bloody Flux, Scowring & Flux	185	Griping in the Guts	1288	Rising of the Lights	397
Burnt and Scalded	8	Hangd & made away themselves	7	Rupture	34
Calenture	3	Headmouldshot & Mouldfallen	14	Scurvy	105
Cancer, Gangrene and Fistula	56	Jaundies	110	Shingles and Swine pox	2
Canker, and Thrush	111	Impostume	227	Sores, Ulcers, broken and bruised Limbs	82
Childbed	625	Kild by severall accidents	46	Spleen	14
Chrisomes and Infants	1258	Kings Evill	86	Spotted Feaver and Purples	1929
Cold and Cough	68	Leprosie	2	Stopping of the stomack	332
Collick and Winde	134	Lethargy	14	Stone and Strangury	98
Consumption and Tissick	4808	Livergrown	20	Surfet	1251
Convulsion and Mother	2036	Meagrom and Headach	12	Teeth and Worms	2614
Distracted	5	Measles	7	Vomiting	51
Dropsie and Timpany	1478	Murthered and Shot	5	Wenn	[illegible]
Drowned	50	Overlaid & Starved	45		

What exactly is statistics ?

- Exploration and description of sample data
- Statistical inference and hypothesis testing
- Stochastic modelling of uncertainty
- Design of experiments and surveys

Basic concepts



- Descriptive statistics – exploration and summary of a sample of data
- Inferential statistics – use of sample data to infer properties of the whole population

Statistical software

- Statistical language-based software

  e.g. Splus, R, SAS
- Interactive Spreadsheet-like packages e.g. Minitab, SPSS, Excel
- Data analysis software with stat routines e.g. MATLAB, PV-Wave, IDL
- Home made subroutines

  e.g. numerical recipes, friend's code, etc.

What exactly is a "model" ?

model **n. [Fr. Modele, It. Modello, from L. modellus]**

A miniature representation (small measure) of a thing, with the several parts in due proportion.

Note 1: **A model is only a "representation" of reality:**
Example 1: Captain Cook's map of Tierra del Fuego.
Example 2: Pablo Picasso's paintings of women's noses

Note 2: **Good modellers know the strong AND weak points of their models**

Note 3: **"Modelling" (English) and "Modeling" (American)**

2. Descriptive sample statistics

1. Data tabulation
2. Summary plots
3. Measures of location, scale, and shape
4. Rank statistics and empirical quantiles
5. Transformation of data values

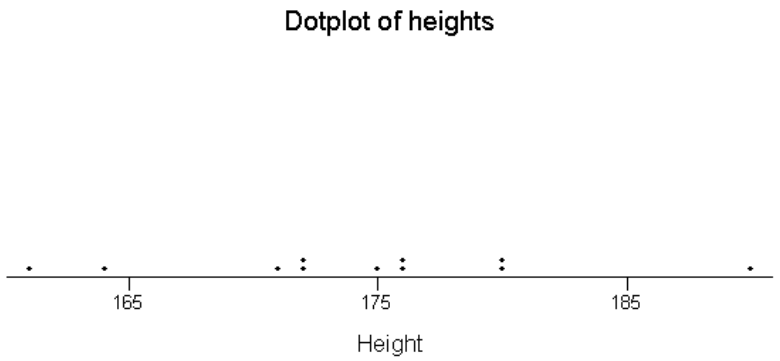
2.1 Data tabulation

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 data

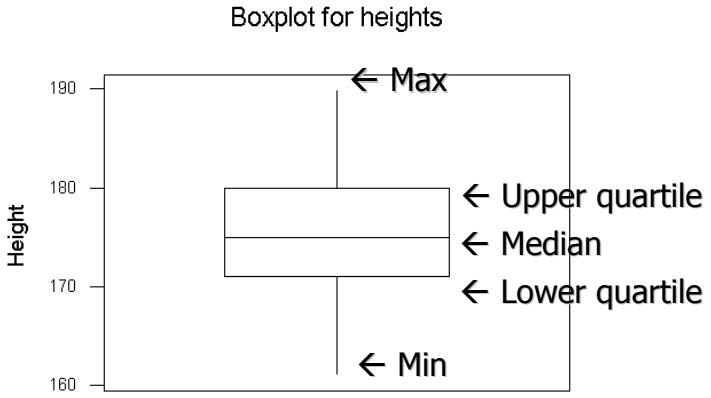
Rows=objects
Columns=variables
Sample size=n=11

2.2a Dotplot (=1-d scatter plot)

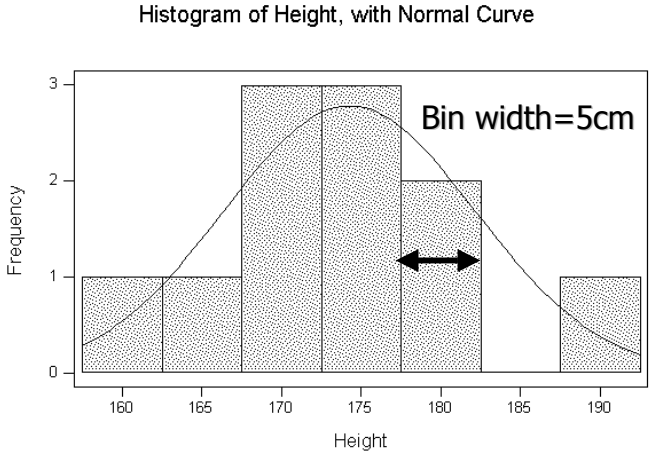


Note the "tied" values that occur in this small sample

2.2c Boxplot

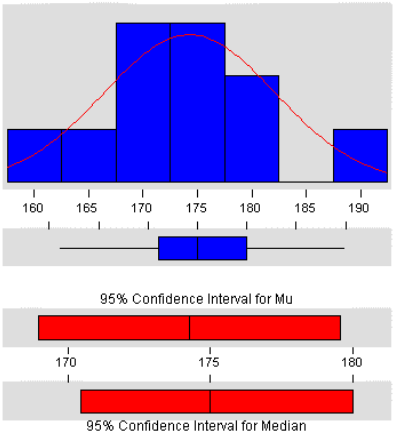


2.2c Histogram



2.2d 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

2.3 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.

2.3 The standard deviation

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

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

2.3 Higher moments about mean

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

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

2.3 Skewness and kurtosis

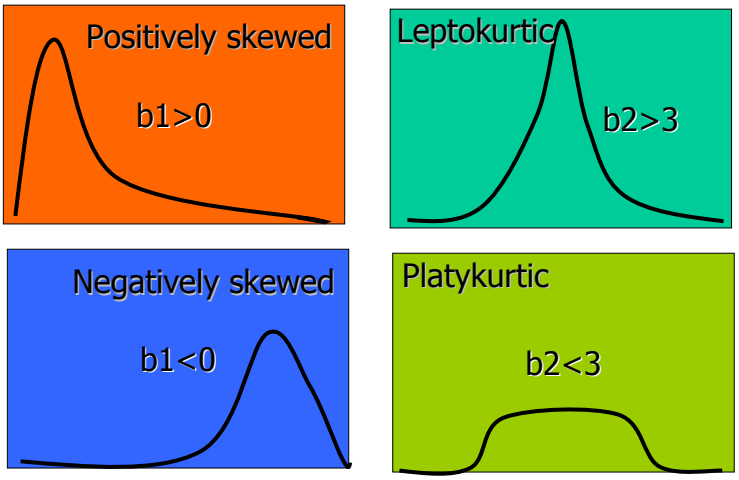
$$Skewness = b_1 = m_3 / s^3$$

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

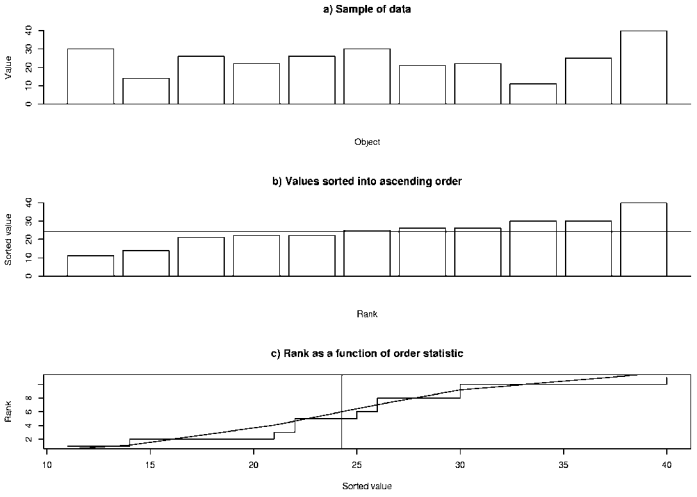
For normal (Gaussian) distribution:
Skewness=0 (symmetric) Kurtosis=3

Kurtosis > 3 "leptokurtic" (fat tails and sharp peak)
Kurtosis < 3 "platykurtic" (thin tails and flatter peak)

2.3 Shapes of distributions



2.4 Empirical 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 data

Rows=objects
Columns=variables
Sample size=n=11

Resistant and Robust statistics

- **Resistant statistic** – not overly sensitive to small or large outlier data
  e.g. IQR compared to max-min range.
- **Robust statistic** – not dependent on the details of the probability distribution
  e.g. rank-statistics (median etc.)

2.5 Transformation of data

- "Center" – remove sample mean
- "Standardize" – remove sample mean and scale
- "Normalize" – nonlinear transformation