Outline

- Definition of Population and Sample in the context of statistics
- Data types and representation
- Numerical summaries of sample properties
- Graphical summaries of sample properties

Learning objectives

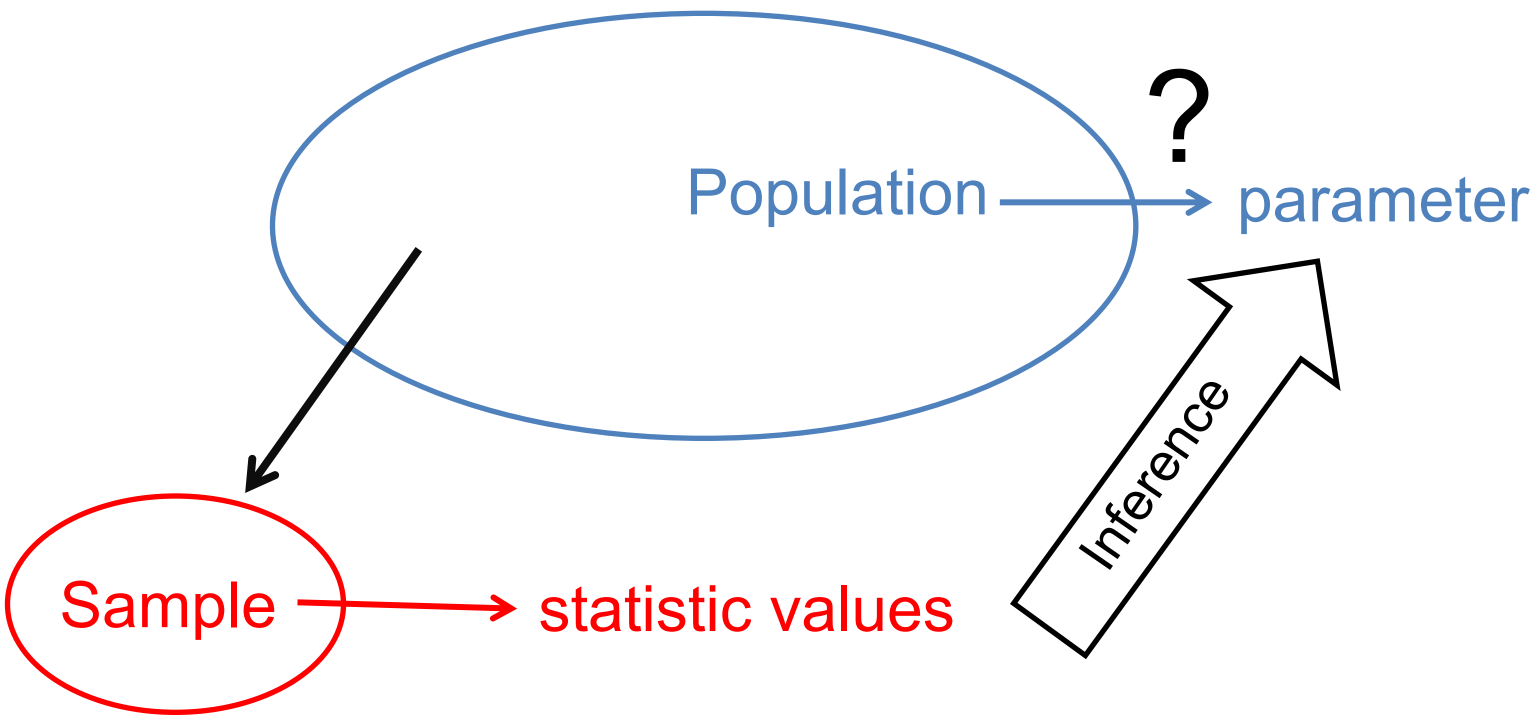
- To learn (or revise?) terminology and fully understand the concepts of population and sample in statistics.

- To recognise different types of variables composing the data (or sample).

- To summarise data and extract information in numerical and graphical form.

Sample versus Population

- Population: The complete set of all possible outcomes in one experiment (making up the entire sample space)
- Sample: A subset of outcomes belonging to a population



Examples - Sample versus Population

- Demographics: The average height of men and women in the UK population can be evaluated from a sample
- Politics: The portion of the population of electors voting for the president of the USA is routinely evaluated from a sample
- Analytical Chemistry: The measurements made to assess the concentration of nitrate ions in a solution is a sample of all the possible measurements (theoretically infinite number) which could be made (which constitute the population).
- Pharmacology: The effect of different drugs on blood pressure is tested in several groups of animals. In this case each group of animals is a sample of the population of all possible animals which could be tested (an infinite number).
- Diagnostics: The diagnostic power of a new MRI method is compared to the existing one in two groups of patients. These groups are samples representative of the diseased population.

Importance of correct sampling

- The *sampling procedure* is critical for the subsequent inference on the population. Inappropriate sampling introduces a systematic error (or bias) on the value to be estimated.

`length, size`

- The *size of the sample* is also important. Intuitively the larger the sample the more accurate the estimate of the population.

(you will see how to determine the sample size required to achieve a specific level of confidence in population estimation)

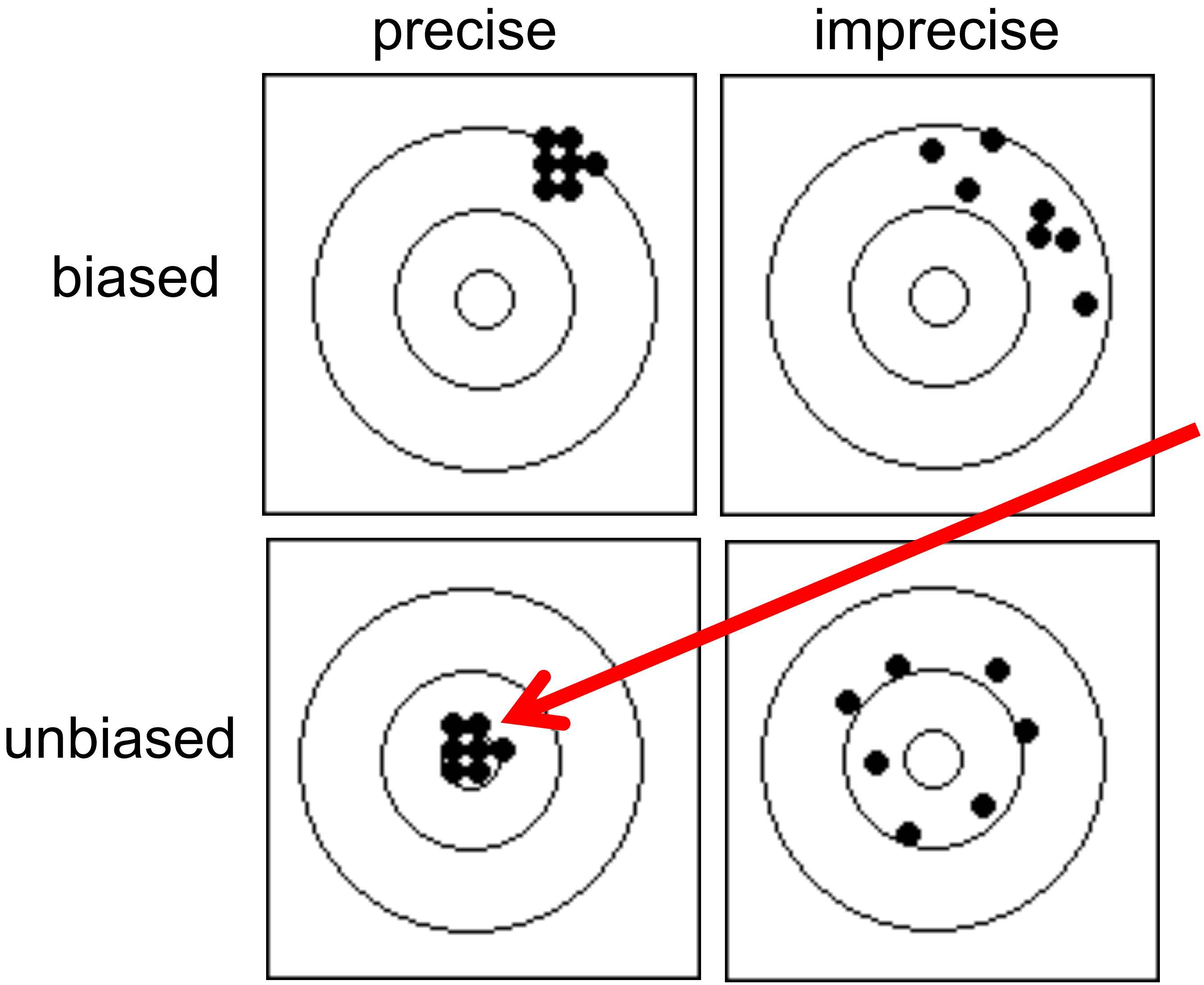
Sampling procedures

- **Random sampling** – most commonly used
  (a set of random numbers is generated and used to select from the whole numbered population)

- **Stratified/Cluster sampling**
  (a set of random numbers is generated and used to select from numbered subsets of the population)

NOTE: Even when sampling the same population, different sampling procedures are likely to lead to different estimates of the same population.

Data distribution versus accuracy of estimation



Which dataset is the most accurate?

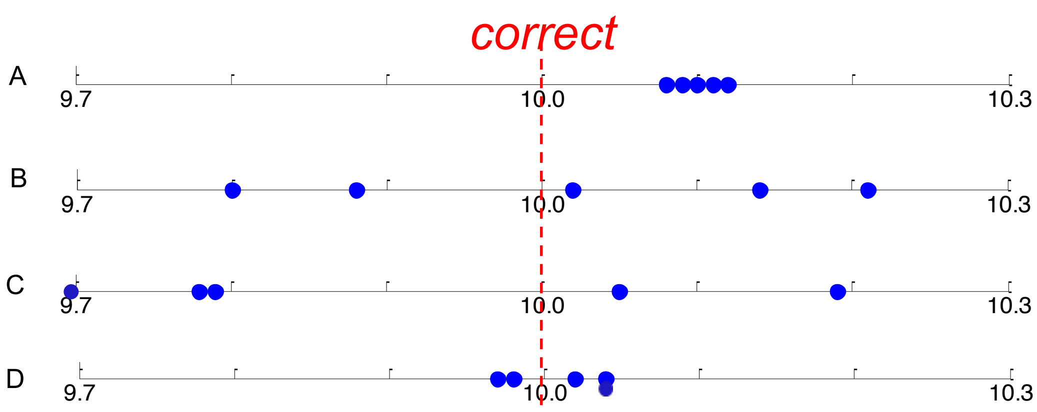
Data distribution versus accuracy of estimation

Example: **Titration measurements**

Each of four students (A,B,C,D) performs an analysis in which exactly 10.00 ml of exactly 0.1 M sodium hydroxide is titrated with exactly 0.1 M hydrochloric acid. Each student performs five replicate titrations, with the results shown in the table below.

Student	**Results (ml)**				
A	10.08	10.11	10.09	10.1	10.12
B	9.88	10.14	10.02	9.8	10.21
C	10.19	9.79	9.69	10.05	9.78
D	10.04	9.98	10.02	9.97	10.04

Accurate one?



Precise, biased

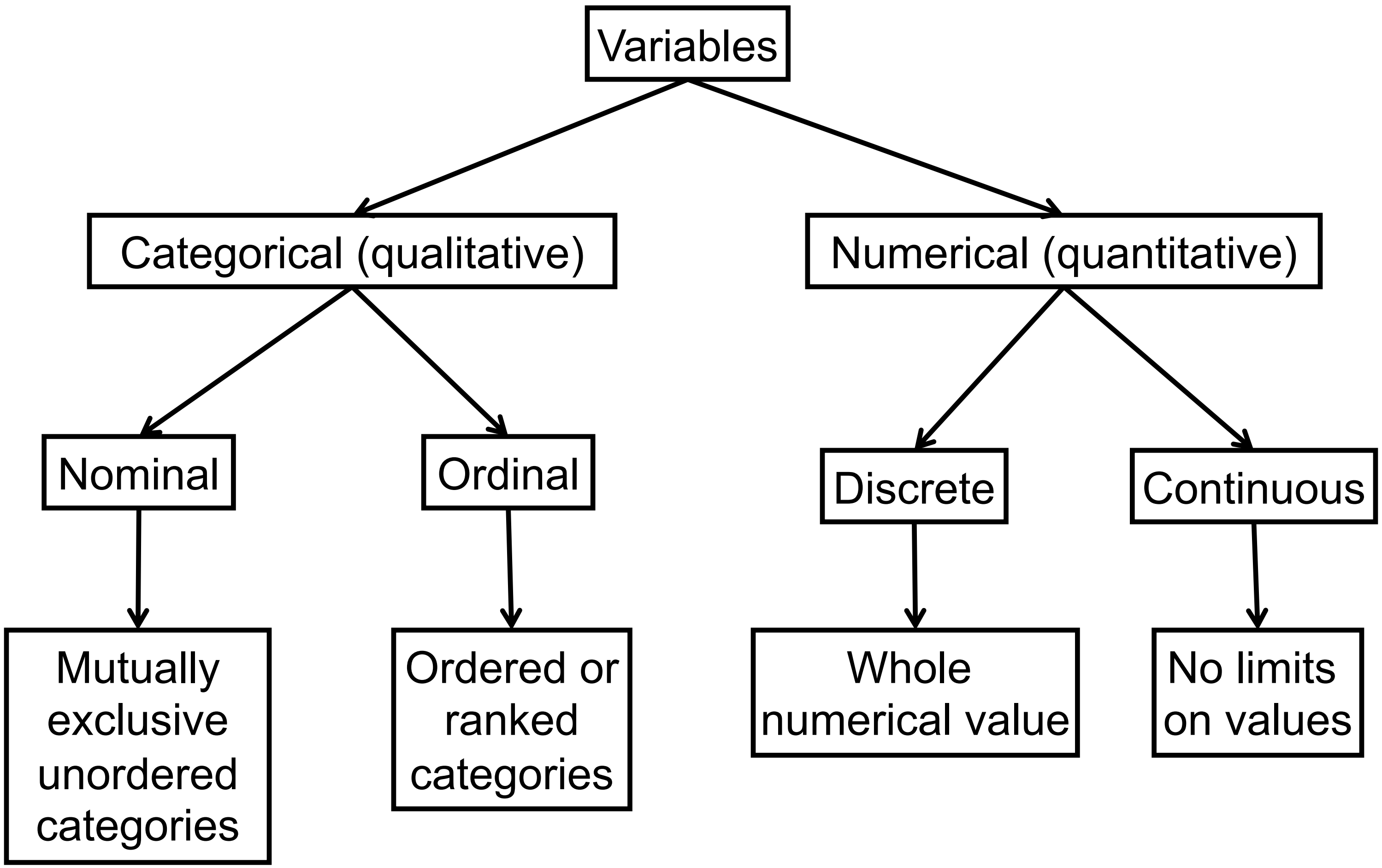
Imprecise, unbiased

Imprecise, biased

Precise, unbiased

Imperial College London

Categorising Data composing the sample



Describing data with numerical measures (1/4)

- Measures of Location are useful for locating the centre of the distribution

mean

  - **Arithmetic mean or average** => sum of the measurements divided by n. Locate the centre if the distribution of values is *symmetric*

population mean: $$\mu = \frac{\sum_{i=1}^{N} X_i}{N}$$ *sample mean:* $$\overline{X} = \frac{\sum_{i=1}^{n} X_i}{n}$$

Describing data with numerical measures (1/4)

- Measures of Location are useful for locating the centre of the distribution

mean

  - **Arithmetic mean or average** => sum of the measurements divided by n. Locate the centre if the distribution of values is *symmetric*

population mean: $$\mu = \frac{\sum_{i=1}^{N} X_i}{N}$$

sample mean: $$\overline{X} = \frac{\sum_{i=1}^{n} X_i}{n}$$

  - **Geometric mean**

$$\sqrt[n]{X_1 \times X_2 \times \cdots \times X_n} = \left(\prod_{i=1}^{n} X_i \right)^{1/n}$$

geomean

  - **Harmonic mean**

$$\frac{n}{1/X_1 + 1/X_2 + \cdots 1/X_n} = \frac{n}{\sum_{i=1}^{n} 1/X_i}$$

harmmean

Describing data with numerical measures (1/4)

- Measures of Location are useful for locating the centre of the distribution

mean

  - **Arithmetic mean or average** => sum of the measurements divided by n. Locate the centre if the distribution of values is *symmetric*

population mean: $$\mu = \frac{\sum_{i=1}^{N} X_i}{N}$$ *sample mean:* $$\overline{X} = \frac{\sum_{i=1}^{n} X_i}{n}$$

  - **Median** of a set of n measurements is the value of x that falls in the middle position when measurements are ordered from smallest to largest.

  Locate the centre when the distribution of values is *asymmetric or "skewed"*

$$Me = \begin{cases} X_{((n+1)/2)} & \text{if n is odd} \\ (X_{(n/2)} + x_{(n/2+1)})/2 & \text{if n is even} \end{cases}$$

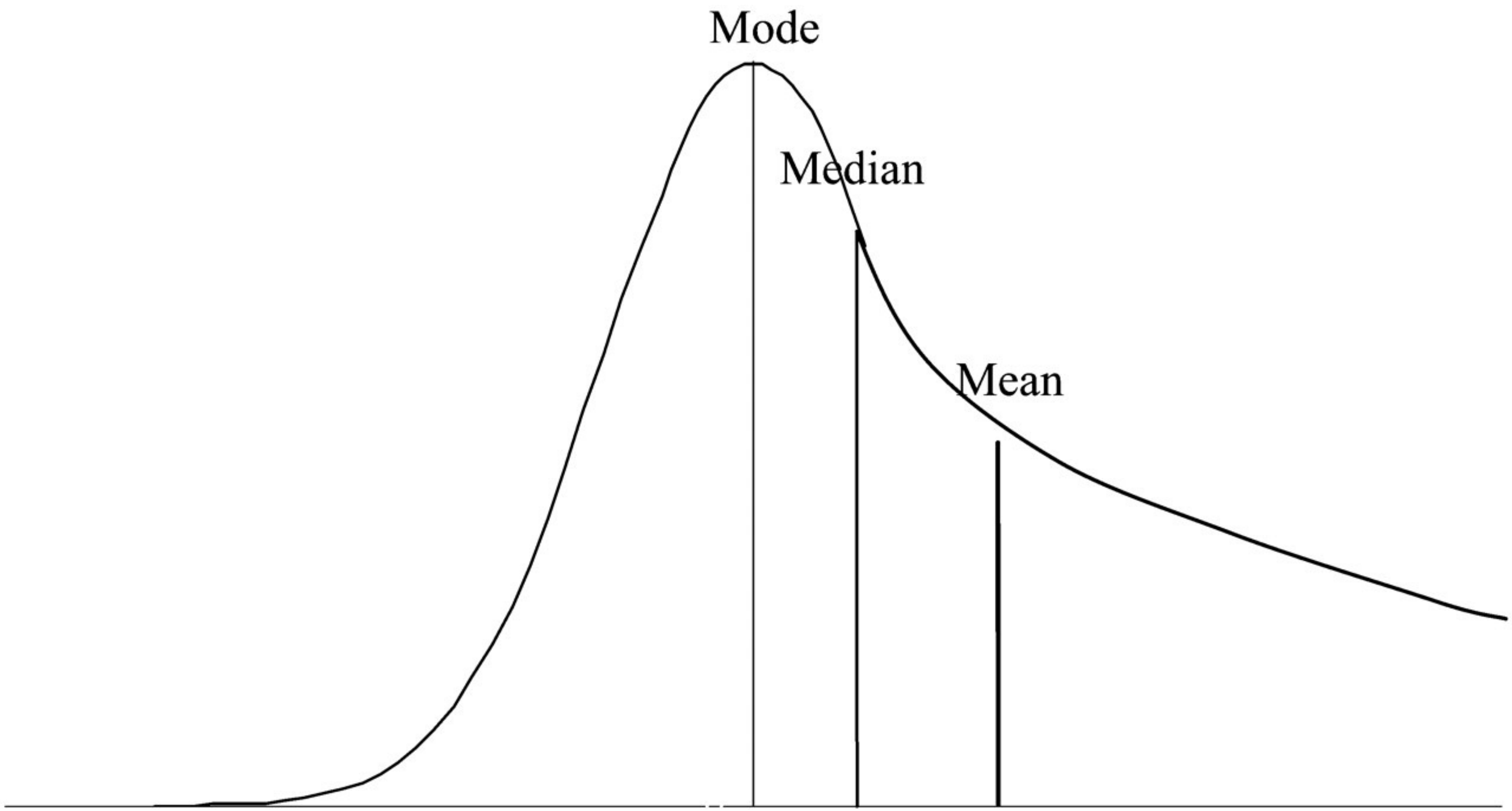
median

  - **Mode** is the most frequently occurring value of x or most frequent category

mode

Describing data distribution with the histogram (3/3)

- Location of the centre of the distribution of the data



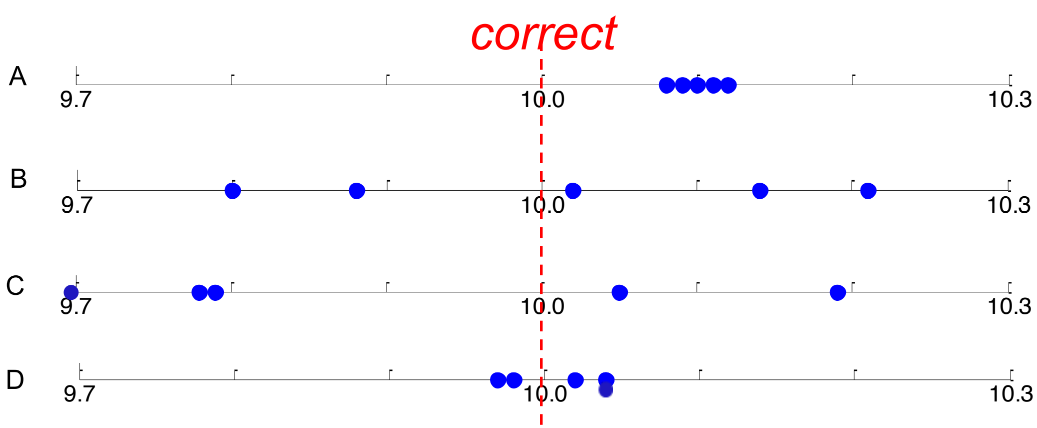
Imperial College
London

Measures of data distribution

Example: **Titration measurements**

Student	Results (ml)				
A	10.08	10.11	10.09	10.1	10.12
B	9.88	10.14	10.02	9.8	10.21
C	10.19	9.79	9.69	10.05	9.78
D	10.04	9.98	10.02	9.97	10.04

Accurate one?



Precise, biased

Imprecise, unbiased

Imprecise, biased

Precise, unbiased

Describing data with numerical measures (2/4)

- Measures of Variability or Dispersion

  – **Range**: difference between the largest and smallest measurements

  `range`

  – **Variance**: average of the sum of squared deviations (i.e. differences between individual measurements x_i and the mean)

$$\sigma^2 = \frac{\sum_{i=1}^{N}(x_i - \mu)^2}{N}$$

σ^2 used for variance of a *population* of N measurements

`var, std`

$$s^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}$$

s^2 used for variance of a *sample* of n measurements

– The **standard deviation (SD)** is the square root of the variance so the variability is expressed in the same units as the sample.

Standard deviation/error?

- Standard Deviation

  – **Variance**: a useful indicator of the dispersion/spread of the data, but the units are the units of x^2 – not so useful!

  – **standard deviation (SD)** is the square root of the variance

$$s^2 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n-1}$$

s^2 is the variance of a *sample* of n measurements

`var, std`

$$S.D. = \sqrt{\frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n-1}}$$

s is the standard deviation of a *sample* of n measurements

We also have the *Standard Error of the Mean,* or *S.E.,* of the *sample*

$$S.E. = \frac{\sigma}{\sqrt{n}}$$

(where σ^2 is the variance of the *population*)

Describing data with numerical measures (3/4)

- Why the sample standard deviation expression differs from population standard deviation ?
  - ie why divide by n-1 rather than n?
  - “if the difference between n and n-1 ever matters to you, you are probably up to no good anyway”

    Numerical Recipes

Describing data with numerical measures (3/4)

- Why the sample standard deviation expression differs from population standard deviation ?
  - **Intuitive answer:** since each x_i tends to be closer to their average $\overline{x}$ than μ , we compensate for this by using the divisor (n-1) rather than n
  - **Theoretical answer:**

    $$\sum_{i=1}^{n}(x_i - \overline{x}) = 0$$

    Only n-1 independent residuals, as they sum to zero

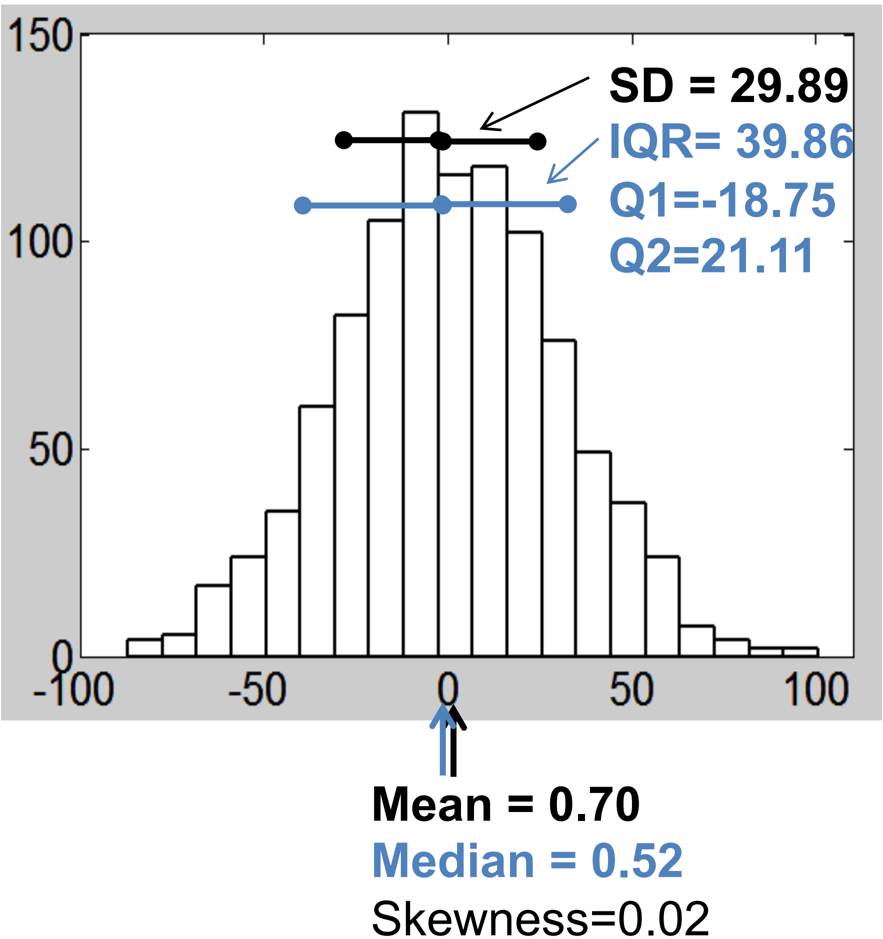
    therefore the nth difference $(x_n - \overline{x})$ can be obtained from the previous (n-1) differences. This information is stored in the sample mean.

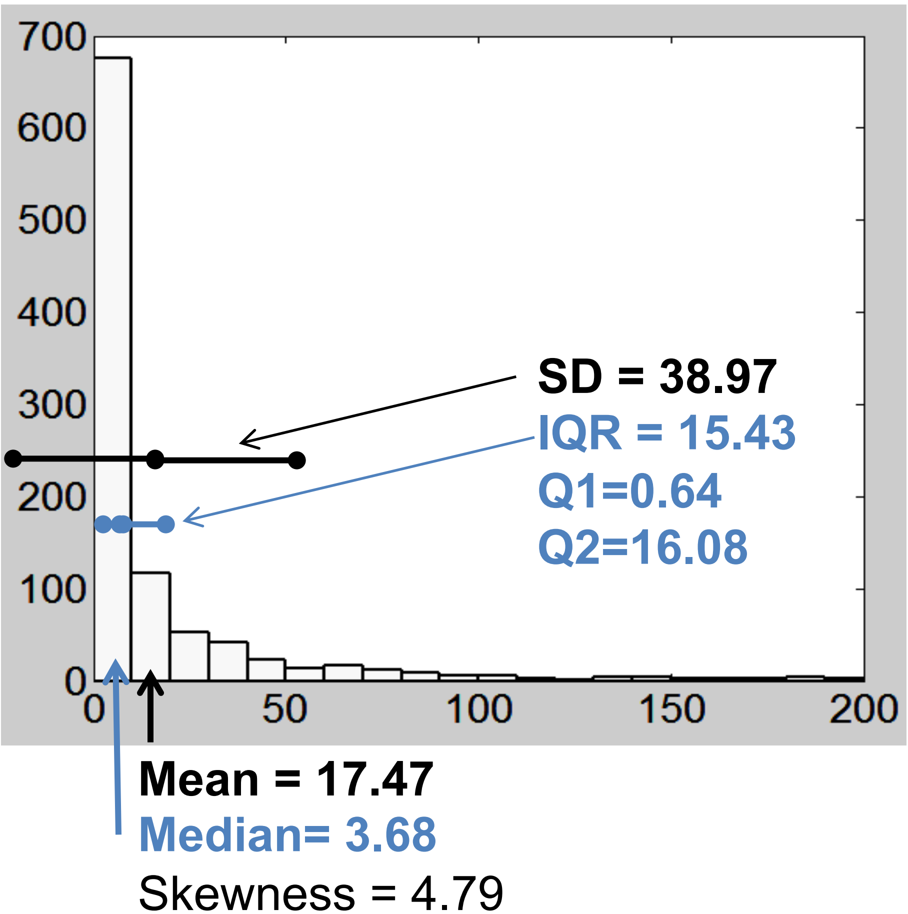
    This is why s is referred as being based on (n-1) “degrees of freedom”.

  - **Empirical answer :** using simulation in Matlab we will take many samples from a population with known σ and show that the sample s is closer to σ when using (n-1) instead of (n) as denominator. Try it…

Describing data with numerical measures (4/4)

- The standard deviation can cause gross mistakes when used to describe the dispersion of the mean in a sample with asymmetric distribution





Describing data and probabilities with graphs

- Describing quantitative data by graphical methods
  - Histograms and bar charts of frequency of occurrence
  - Stem and Leaf plots
  - Pie Charts
  - Box and Whiskers plots
  - Scatter plots and Line plots
  - Empirical Cumulative Distribution Function (ECDF)
  - Q-Q plots

(Note: Implementation of graphical methods in Matlab will done in the lab sessions)

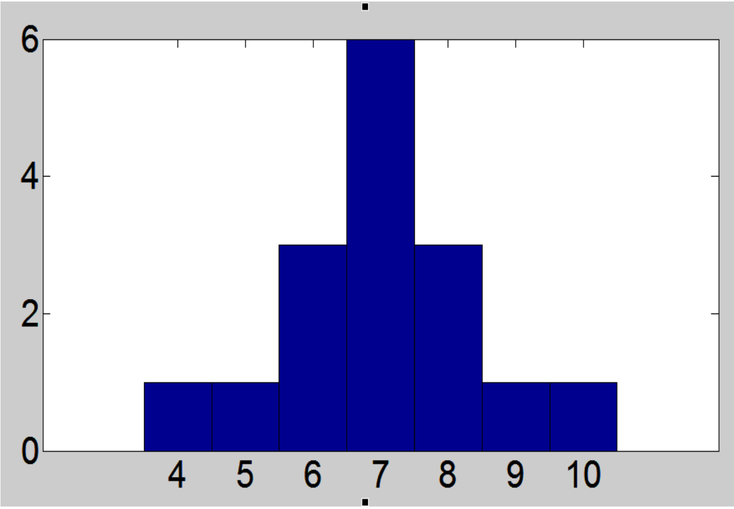
- Describing probabilities by tree diagrams and probability tables

Describing data distribution with the histogram (1/3)

- **Histogram** `hist, histc`

The primary way of summarising variability is via the frequency distribution.

Frequency= Number of times an event has happened



An histogram plots **frequency** for **interval-grouped** data.

Consider data set: 4, 5, 6, 6, 6, 7, 7, 7, 7, 7, 7, 8, 8, 8, 9, 10

Frequency table

4	1
5	1
6	3
7	6
8	3
9	1
10	1

This leads to the symmetric histogram in the figure.

Note: Median = Mean = 7

Describing data distribution with the histogram (2/3)

- **Asymmetric histogram (Skewed distributions)**

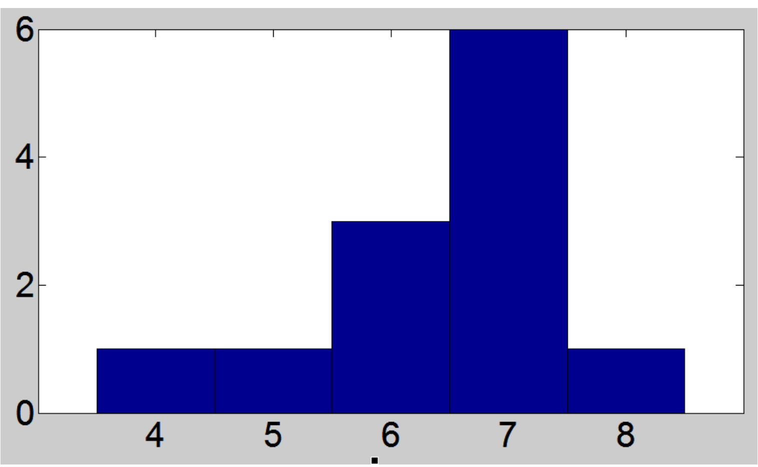
A. Consider the previous data set but truncated on the right:

4, 5, 6, 6, 6, 7, 7, 7, 7, 7, 7, 8.

B. Then truncated on the left:

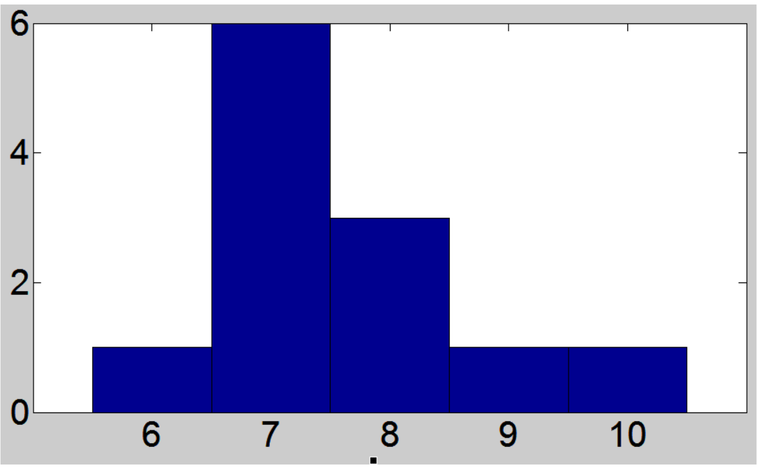
6, 7, 7, 7, 7, 7, 7, 8, 8, 8, 9, 10

A.



Skewed to the left
Median=7 > Mean=6.4

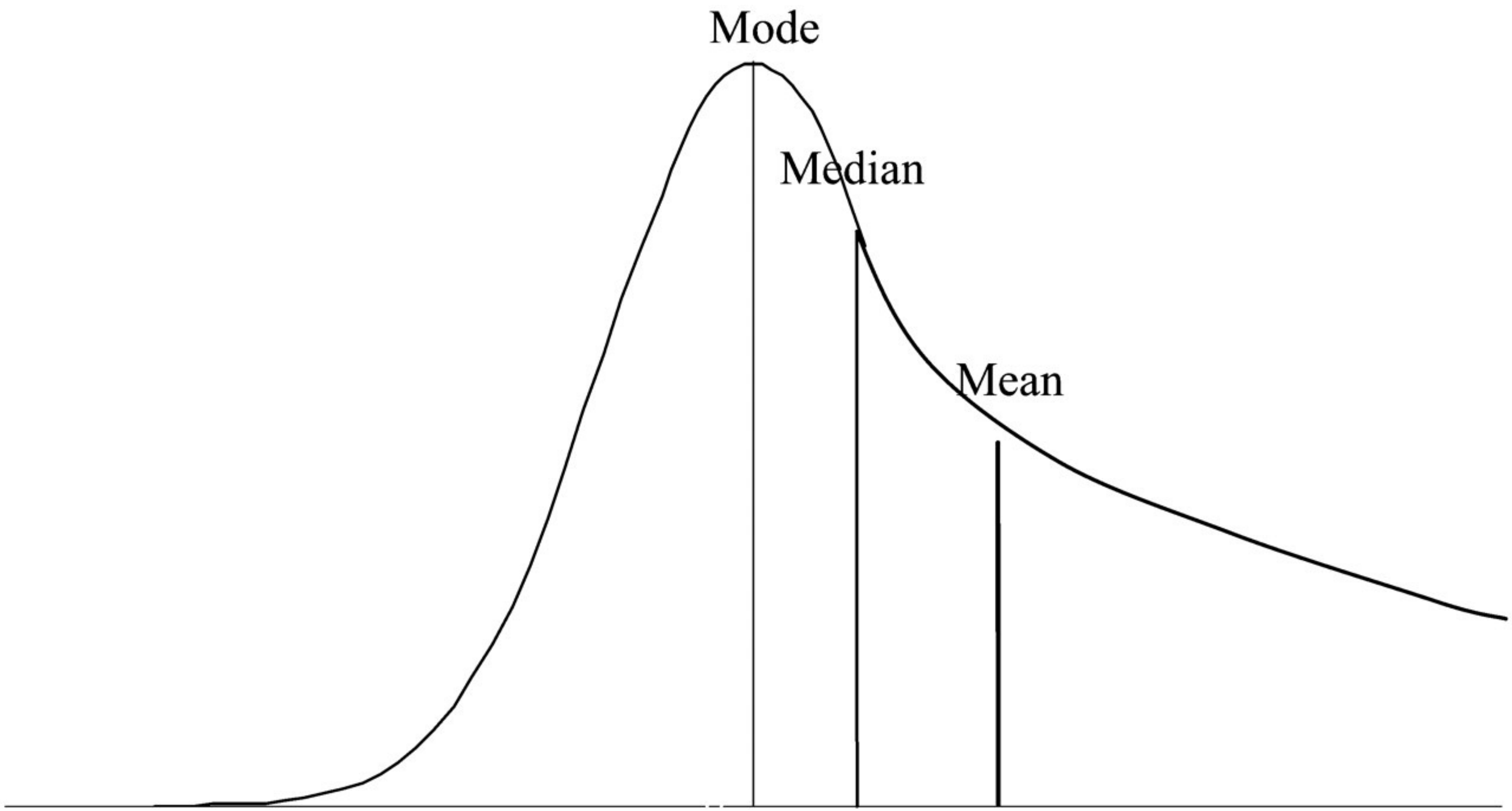
B.



Skewed to the right
Median=7 < Mean=7.6

Describing data distribution with the histogram (3/3)

- Location of the centre of the distribution of the data



Additional numerical descriptors

Moments of higher order

$$m_k = \frac{1}{n}\sum_{i=1}^{n} X_i^k$$

raw sample moments (e.g. m_1 = mean)

$$\mu_k = \frac{1}{n}\sum_{i=1}^{n} (X_i - m_1)^k$$

Central moments of order k (e.g. μ_2 = variance)

- **Skewness**: measures the degree of asymmetry in a sample distribution

`skewness`

$$\gamma_n = \mu_3 / \mu_2^{3/2} = \mu_3 / s_*^3$$

$\gamma_n > 0$ => right tail => skewed to the right
$\gamma_n < 0$ => left tail => skewed to the left

- **Kurtosis**: measure the "peakedness" or flatness of a sample distribution

`kurtosis`

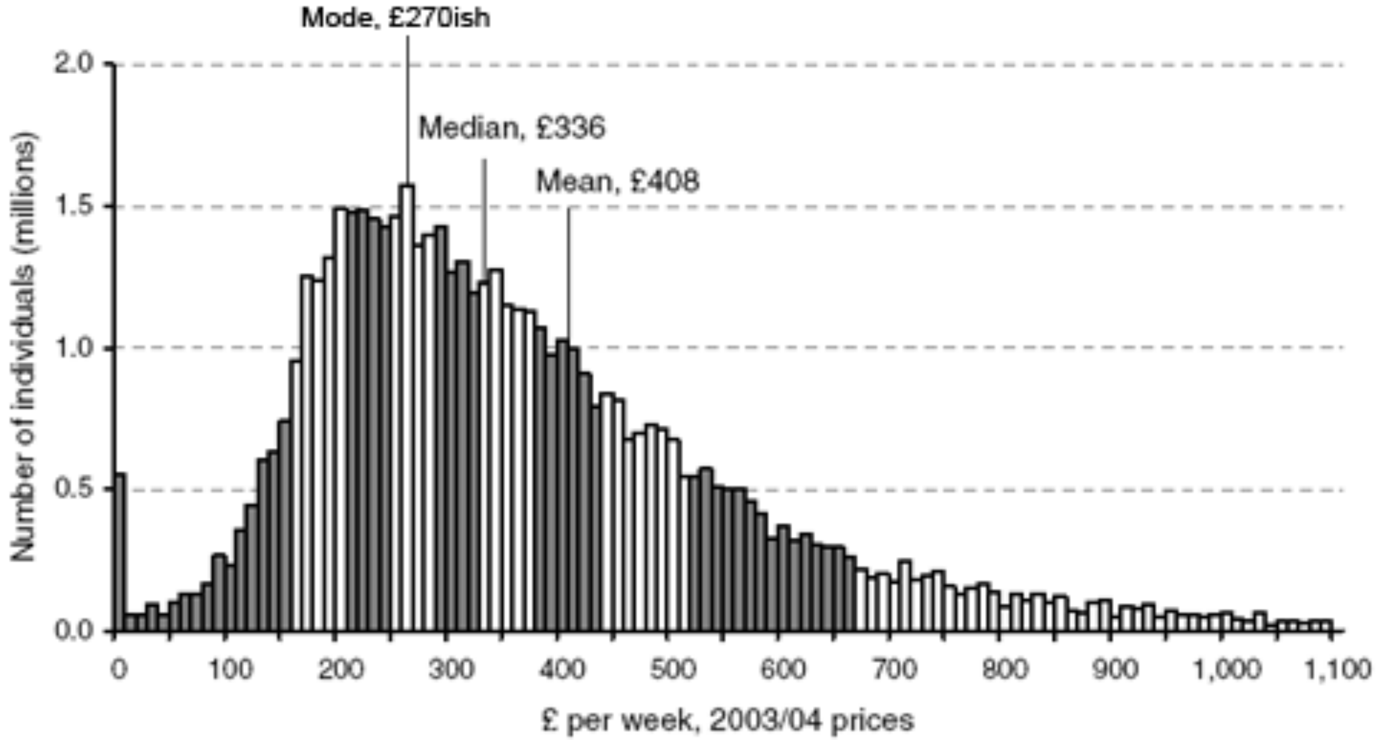
$$\kappa_n = \mu_4 / \mu_2^2 = \mu_4 / s_*^4$$

$\kappa_n = 3$ reference shape (e.g. gaussian)
$\kappa_n > 3$ => more peaked (*leptokurtic*)
$\kappa_n < 3$ => flatter (*platykurtic*)

Note: Where s_* = is the sample SD calculated with (n) instead of (n-1) as denominator

Skewed?

Figure 2.1. The income distribution in 2003/04



Skewness & Kurtosis

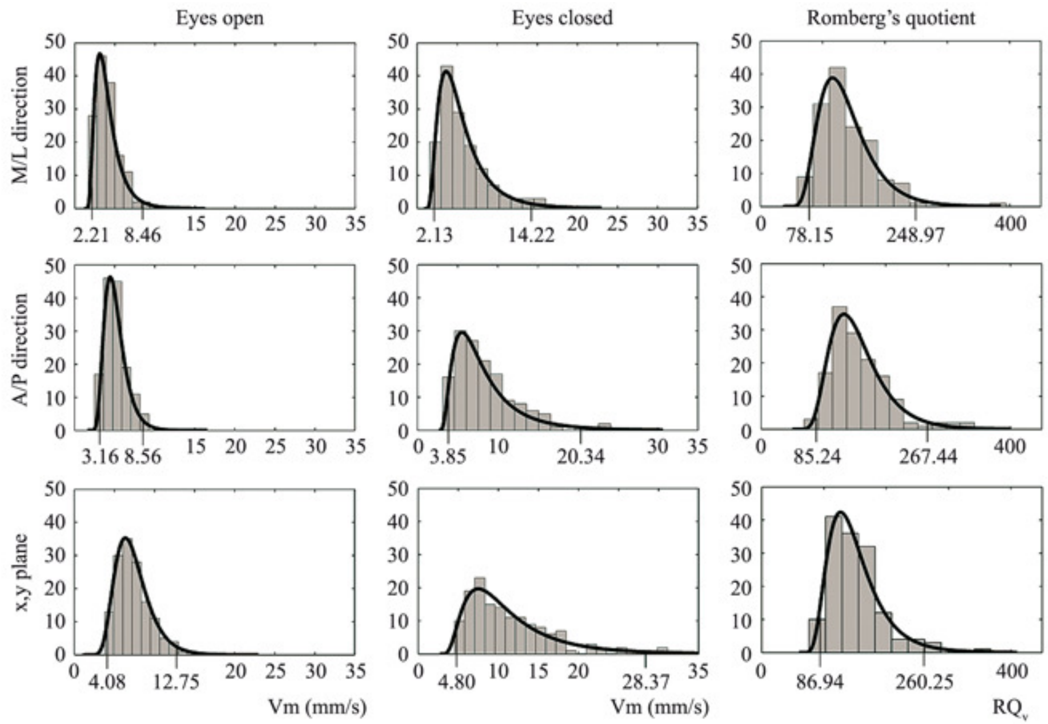


Figure 2. The histograms and the lognormal curves fitted to the experimental data. The values detached indicate the threshold scores for the 0.025 to 0.975 confidence levels.

Describing data and probabilities with graphs (1/7)

- ## Stem and Leaf Plot

no built-in function

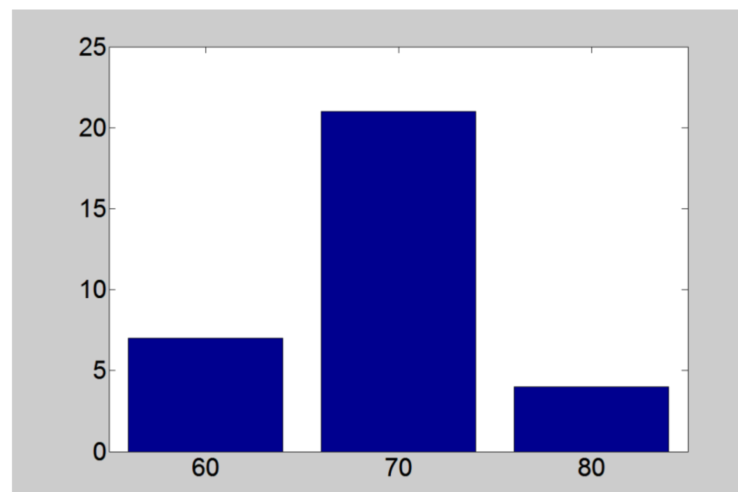
Method to display data in a structured list.

Example: Tibetan skull height dataset – what is the distribution?

[74 63 70 65 78 72 71 74 70 62 71 65 75 77 68 71 66 76 74 73 77 79 72 80 77 76 83 82 74 79 81 69]

6	2 3 5 5 6 8 9
7	0 0 1 1 1 2 2 3 4 4 4 4 5 6 6 7 7 7 8 9 9
8	0 1 2 3

Equivalent histogram



Describing data and probabilities with graphs (2/7)

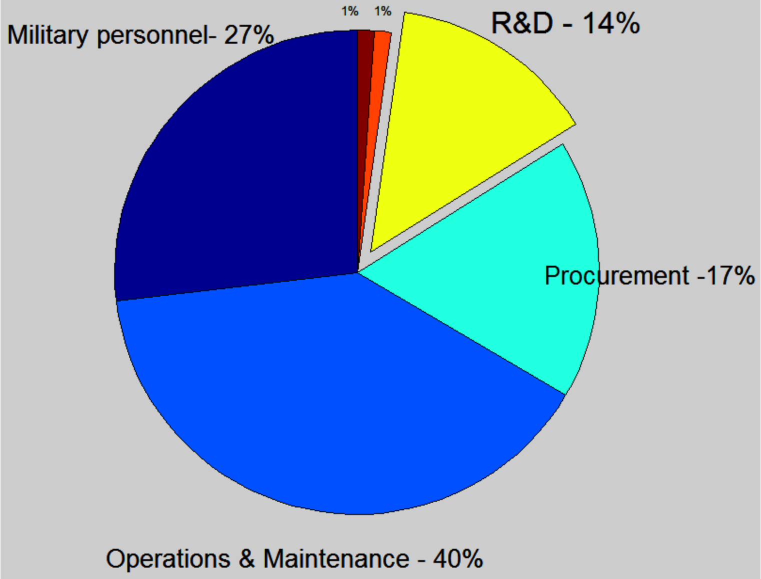
- Pie Charts

`pie, pie3`

Appropriate for visualising proportions or frequencies

Example – visualising the proportion of budget expenditure by US Department of Defence in fiscal year 2005

Category	Amount (billions USD)
Military personnel	127.5
Operations & Maintenance	188.1
Procurement	82.3
R&D	65.7
Military construction	5.3
Other	5.5
Total	474.4



Describing data and probabilities with graphs (3/7)

- Box and Whiskers plots

`boxplot, quantile`

Data	1	11.5	6	7.2	4	8	9	10	6.8	8.3	2	2	10	1
Sorted	1	1	2	2	4	6	6.8	7.2	8	8.3	9	10	10	11.5

Me=7

Quartiles: represent the spread of a data set by breaking the data set into quarters.
Q2 = Me = 7
Q1= data point located at 25% of the ordered data set (25% quantile)
Q3= data point located at 75% of the ordered data set (75% quantile)

How are quantile calculated? There are different approaches.

In Matlab: the ordered values in the dataset are taken as:
(0.5/n), 1.5/n),……,([n-0.5]/n) (here multiplied by 100 to show Q1 and Q3)

3.6	10.7	17.8	25	32.1	39.3	46.4	53.6	60.7	67.8	75	82.1	89.3	96.4

Describing data and probabilities with graphs (4/7)

- Box and Whiskers plots

`boxplot, quantile`

Data	1	11.5	6	7.2	4	8	9	10	6.8	8.3	2	2	10	1
Sorted	1	1	2	**2**	4	6	6.8	7.2	8	8.3	**9**	10	10	11.5

$\underbrace{6.8 \quad 7.2}_{Me=7}$

Quartiles: Spread of a data set by breaking the data set into quarters.

Q2 = Me = 7

Q1= 2

Q3 = 9

IQR = Q3 – Q1 = 7

Whiskers:

Lower Threshold = Q1 - 1.5 x IQR

Upper Threshold = Q3 + 1.5 x IQR

Outliers: data outside whiskers

or highest adjacent value

?

UT=11.5

Q3=9

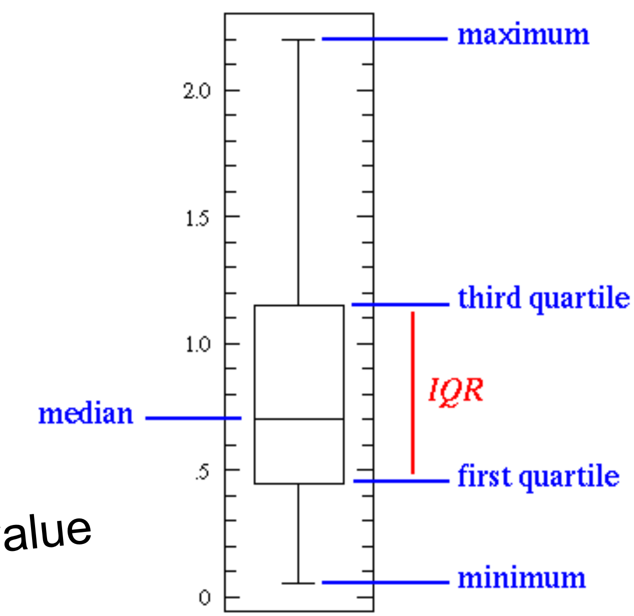
Me=7

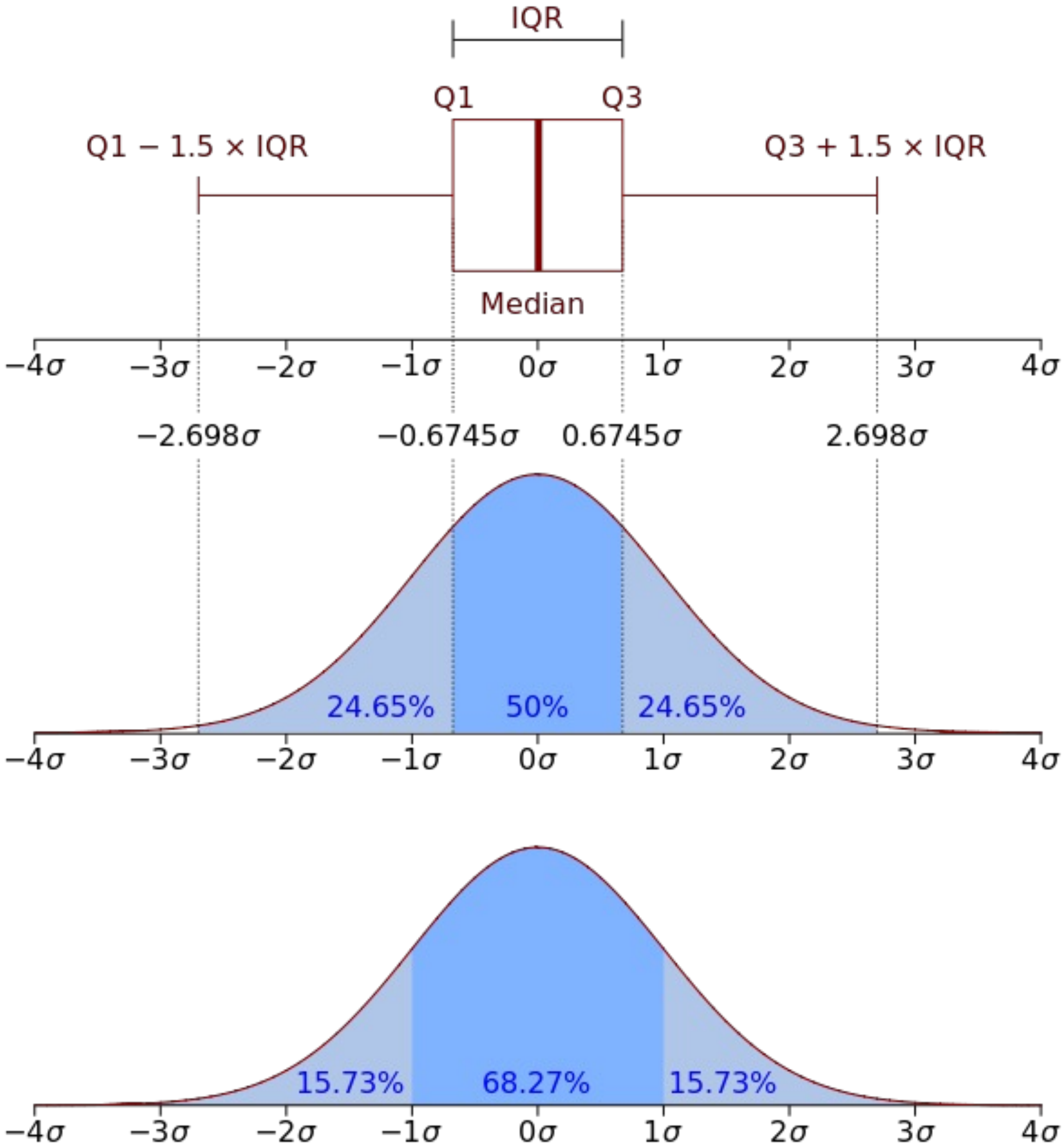
or lowest adjacent value

?

Q1=2

LT=1




IQR
Q1
Q3
Q1 − 1.5 × IQR
Q3 + 1.5 × IQR
Median
−4σ −3σ −2σ −1σ 0σ 1σ 2σ 3σ 4σ
−2.698σ
−0.6745σ
0.6745σ
2.698σ
24.65%
50%
24.65%
−4σ −3σ −2σ −1σ 0σ 1σ 2σ 3σ 4σ
15.73%
68.27%
15.73%
−4σ −3σ −2σ −1σ 0σ 1σ 2σ 3σ 4σ

5 minute test question

What is the median?

175 190 250 230 240 260

200 185 190 195 225 265

175 185 190 190 195 200 225 230 240 250 260 265

Describing data and probabilities with graphs (5/7)

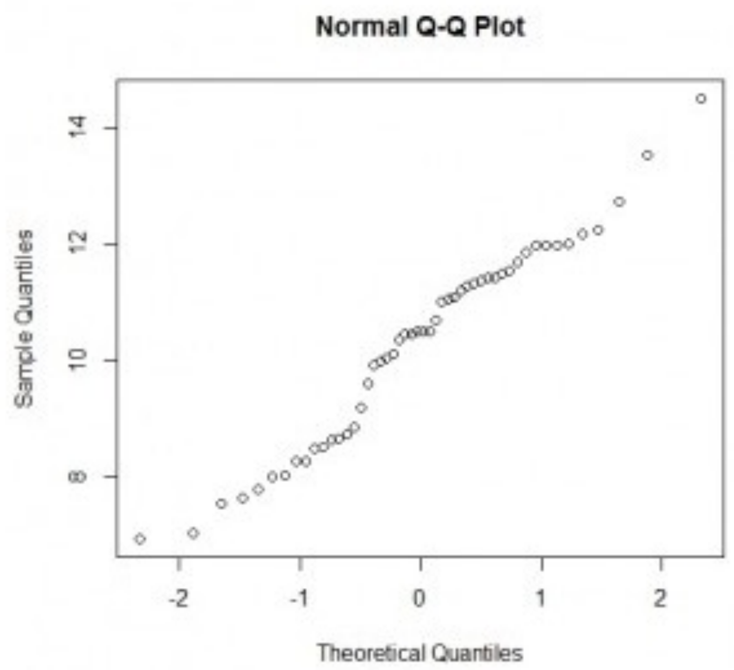
- Q-Q plot

`qqplot, quantile`

The quantile-quantile plots compare the distribution of a sample with the distribution of another sample or with a standard theoretical distribution.

This is done by plotting the sample quantiles of one distribution against the corresponding quantiles of the other.

If the plot is close to linear, then the distributions are close (up to a scale shift).
45° slope => equal distributions



Describing data and probabilities with graphs (6/7)

ecdf

- Empirical Cumulative Distribution Function (ECDF)

ECDF is the accumulation of the previous frequencies, i.e. adding all previous frequencies to the frequency of the current value.

$$F(x) = \frac{1}{n}\sum_{i=1}^{n}\mathbf{1}(X \le x)$$

Not a great notation in Vidakovic…

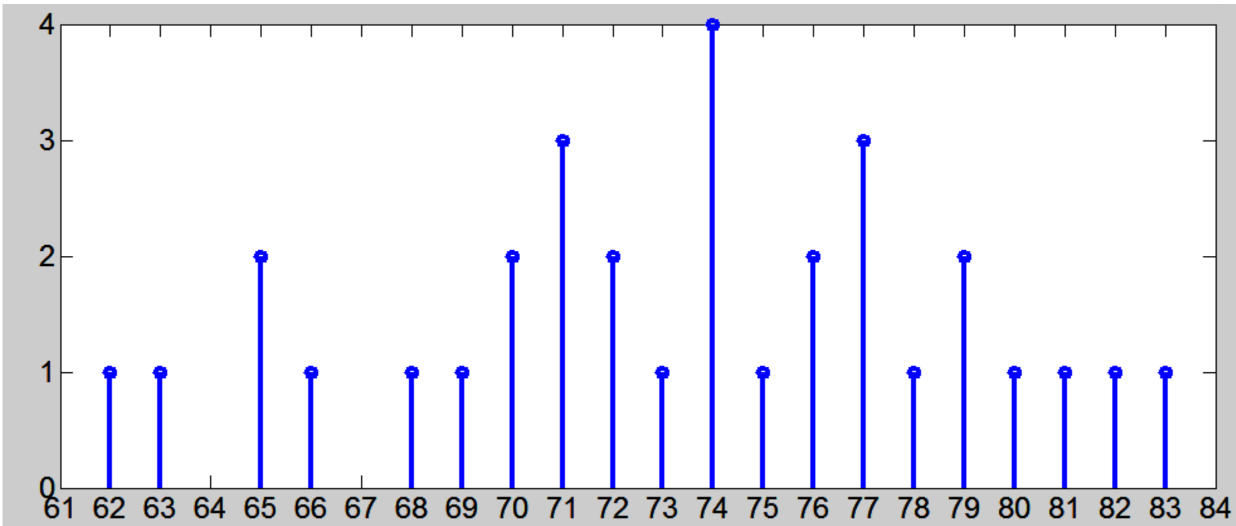
$\mathbf{1}(X \le x)$ is an operator which return 1 if the expression is true and 0 otherwise.

Alternatively consider summing the k frequencies f_i where k is such that $X_k \le x$

[74 63 70 65 78 72 71 74 70 62 71 65 75 77 68 71 66 76 74 73 77 79 72 80 77 76 83 82 74 79 81 69]

- ECDF Example*: Sample of Tibetan skull height (slide 20)*

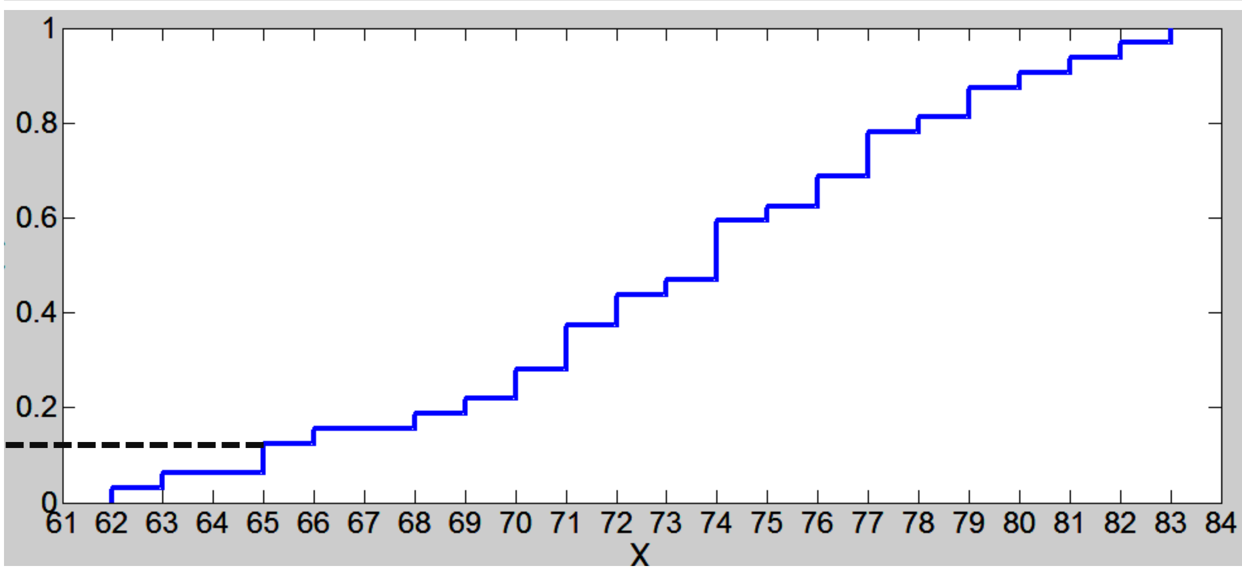
Freq



n= 32

F(x)

0.1250



$F(65) =$
$= 1/32 + 1/32 + 0 +$
$+ 2/32 = 0.125$

Correlation (1/2)

`cov, corr`

- **Correlation in paired samples**

$$X = (X_1, X_2, \dots, X_n)$$
$$Y = (Y_1, Y_2, \dots, Y_n)$$

Sample correlation coefficient ***r*** measures the strength and direction of the linear relationship between two paired samples

$$r = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum_{i=1}^{n}(X_i - \bar{X})^2 \cdot \sum_{i=1}^{n}(Y_i - \bar{Y})^2}}$$

$$Cov(X,Y) = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})$$

$$r = \frac{Cov(X,Y)}{s_X s_Y}$$

-1 < r < 1

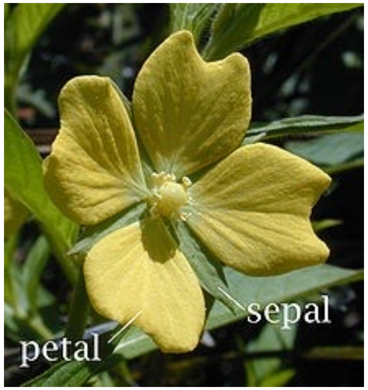
r = 0 => no correlation
| r | ~ 1 => strong correlation

Pearson Correlation Coefficient

Imperial College London

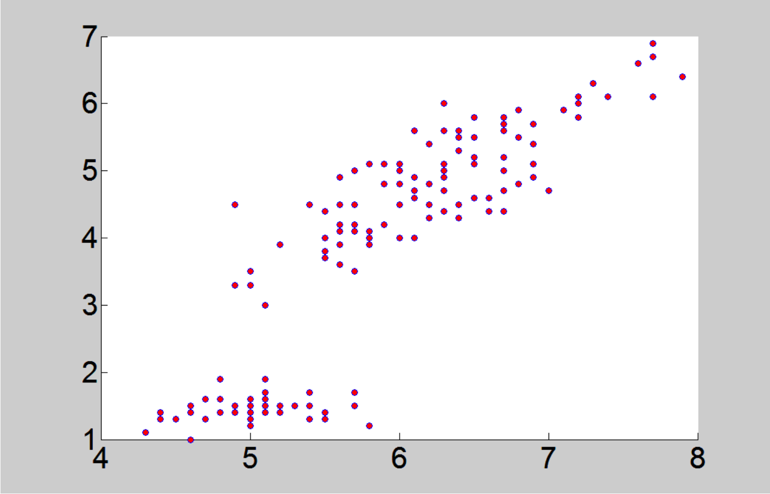
Correlation (2/2)

- Correlation in paired samples



Example: correlation between sepal and petal length in flowering Plants (see Fisher 's iris dataset).

scatter



Cov = 1.2743

not a good indicator of the relationship since it is scale(magnitude) dependent

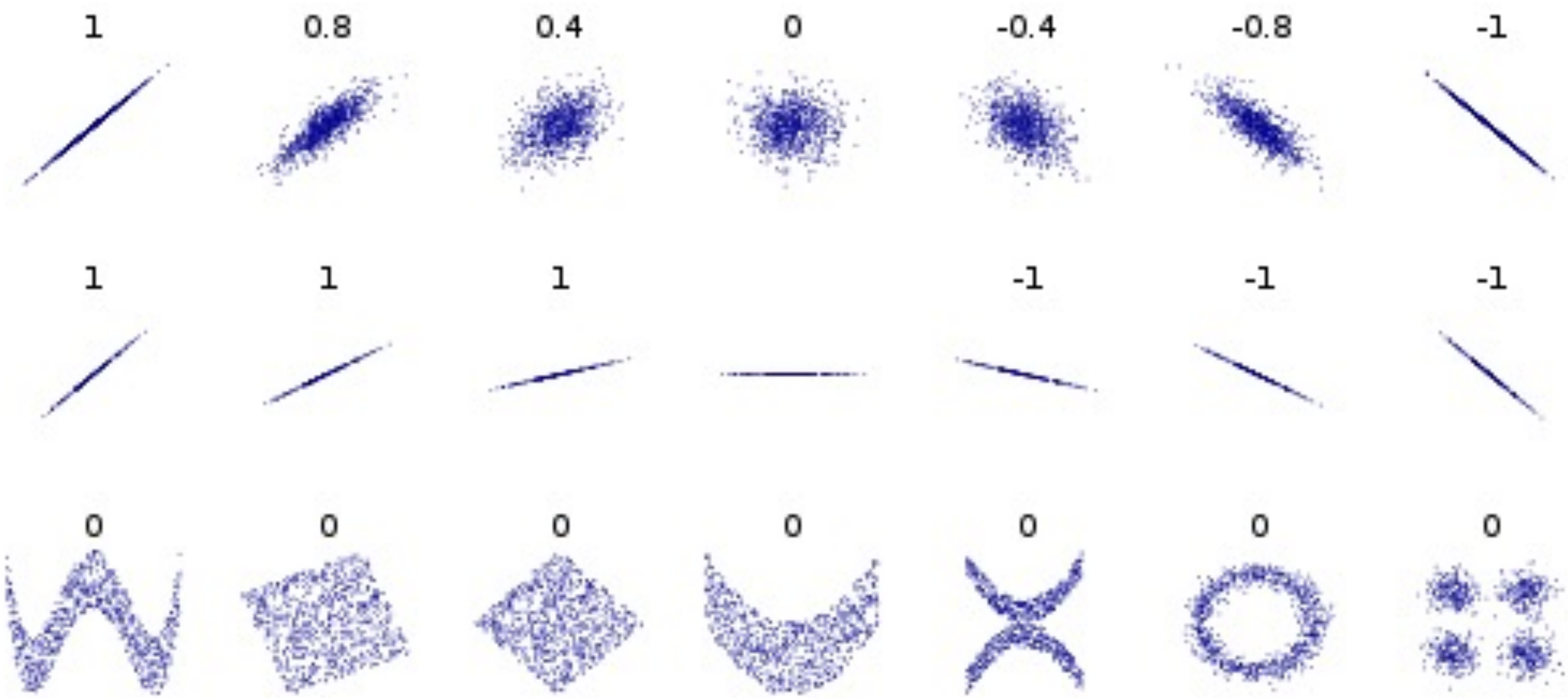
cov

Correlation = r = 0.8718

corr

(Column 1 and 3 in dataset stored in file fisheries.mat in BlackBoard)

Some Pearson correlation coefficients



Lecture 1 References (1/2)

1. **Lecture Slides** in Blackboard (https://bb.imperial.ac.uk) in the BE9-MSTDA -> Course Content -> “Lectures” folder.

2. Statistics for Bioengineering Sciences (with Matlab and WinBUGS support). **B. Vidakovic**, Springer 2011.

   Chapter 2, p.9-42

View only : http://books.google.co.uk/books?id=_HiSXTwpgNgC

Lecture 1 References (2/2)

3. Introduction to statistical thinking (With R, Without Calculus). **Bejamin Yakir**, 2011, The Hebrew University, Israel.

Chapter 1, p.3-6 / Chapter 2, p.15-18 / Chapter 3, p.29-38

Download from web:
http://pluto.huji.ac.il/~msby/StatThink/IntroStat.pdf