MATH1019 Linear Algebra and Statistics for Engineers

Lecture 1: Data Handling

Overview: We look at some basic ways of how data can be summarised and presented both numerically and graphically.

Motivation: Statistics is used in all phases of engineering work relating to research, development, or production. Engineers routinely collect, process, analyse, and interpret numerical data. Data may arise either from a designed experiment or from an observational study. The use of statistics in interpreting this data plays a major role in quality control to improve any engineering process or product. When performing statistical analysis on a set of data, the mean, median, mode, and standard deviation are all helpful values to calculate.

Learning outcomes

In today's lecture we will learn how to:

- Present data using various graphical means
- Summarise data using descriptive statistics

Key concepts in this lecture:

- Overview of Statistics
- Presenting data using Stem-and-Leaf display
- Presenting data using a histogram
- Measures of central tendency
- Measures of dispersion
- Five-number summary
- Box plot and outliers

Introducing R

Laboratory 1

1 Introduction

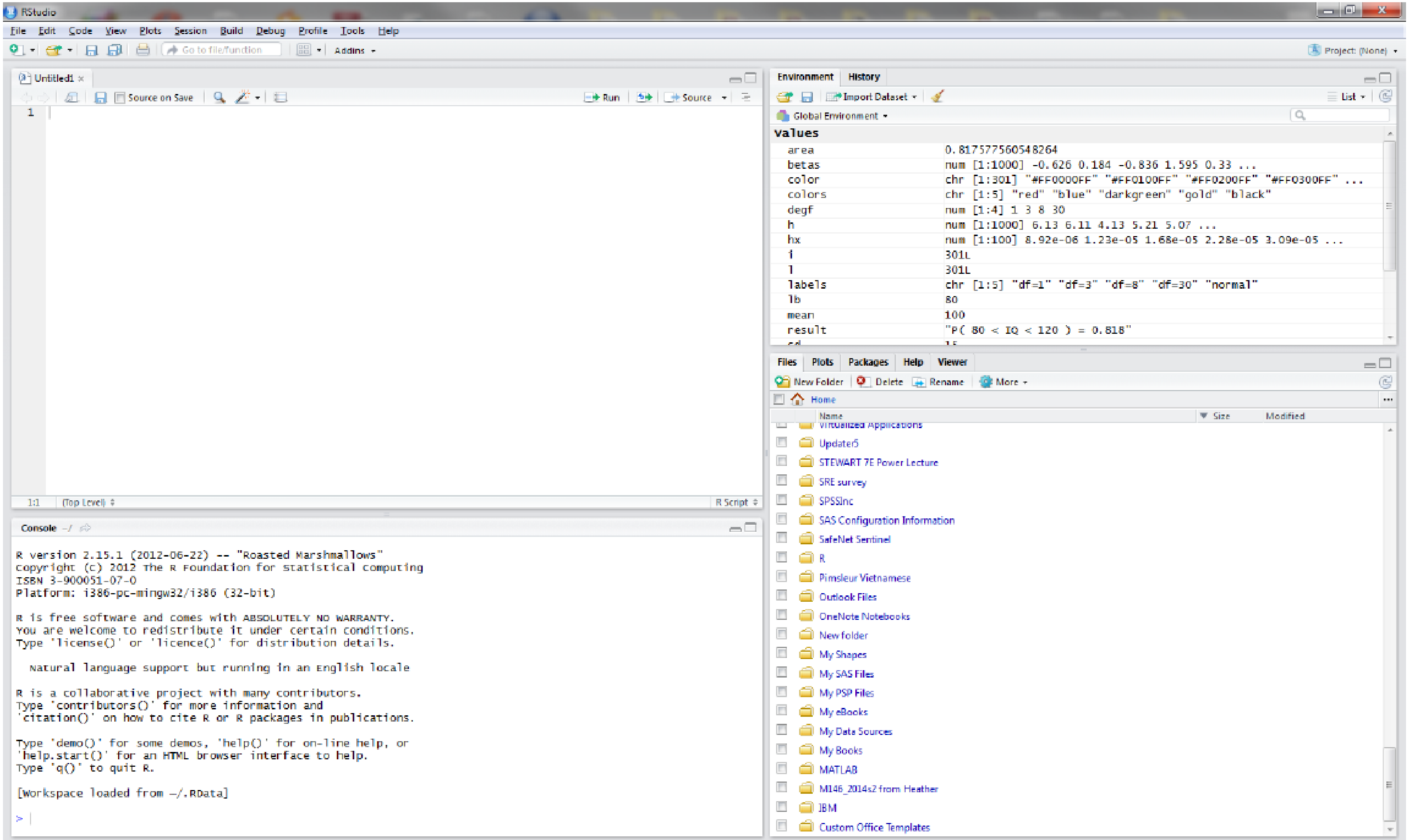
R provides an extremely powerful environment in which you can perform statistical analysis and produce graphics. It is actually a complete programming language. *R* is one of the most widely used statistics programming languages among data scientists and is supported by a vibrant community of contributors.You can use it for simple purposes or very complex ones. In this tutorial, we will only cover some of the basics of *R*.

R is a command driven language, so to make our lives easier we will be using an interface to *R* called *RStudio*. *R* can be freely downloaded from: https://www.r-project.org/ and *RStudio* can be downloaded from: https://www.rstudio.com/ .

2 Getting started

Before you begin, you should create a working directory where all your data files will be stored. When you log in to a Curtin Computer, you will have access to your own drive, the I-drive. OK. We are ready to begin! When you open *RSudio*, you will see a window similar to the following:

Introducing R



What is Statistics?

The science of collecting, describing and interpreting data

Descriptive Statistics

- Collection, presentation, and description of sample data

Inferential Statistics

- Interpreting the values resulting from descriptive techniques and drawing conclusions about a population

Why Study Statistics?

Answers provided by statistical analysis can provide the basis for making decisions or choosing actions. Statistical reasoning and methods can help you become efficient at obtaining information and making useful conclusions.

Example 1

A civil engineer must determine the strength of supports for generators at a power plant. A number of those available must be loaded to failure, and their strengths will provide the basis for assessing the strength of other supports not subject to testing. The proportion of all supports available with strengths that lie below a design limit needs to be determined.

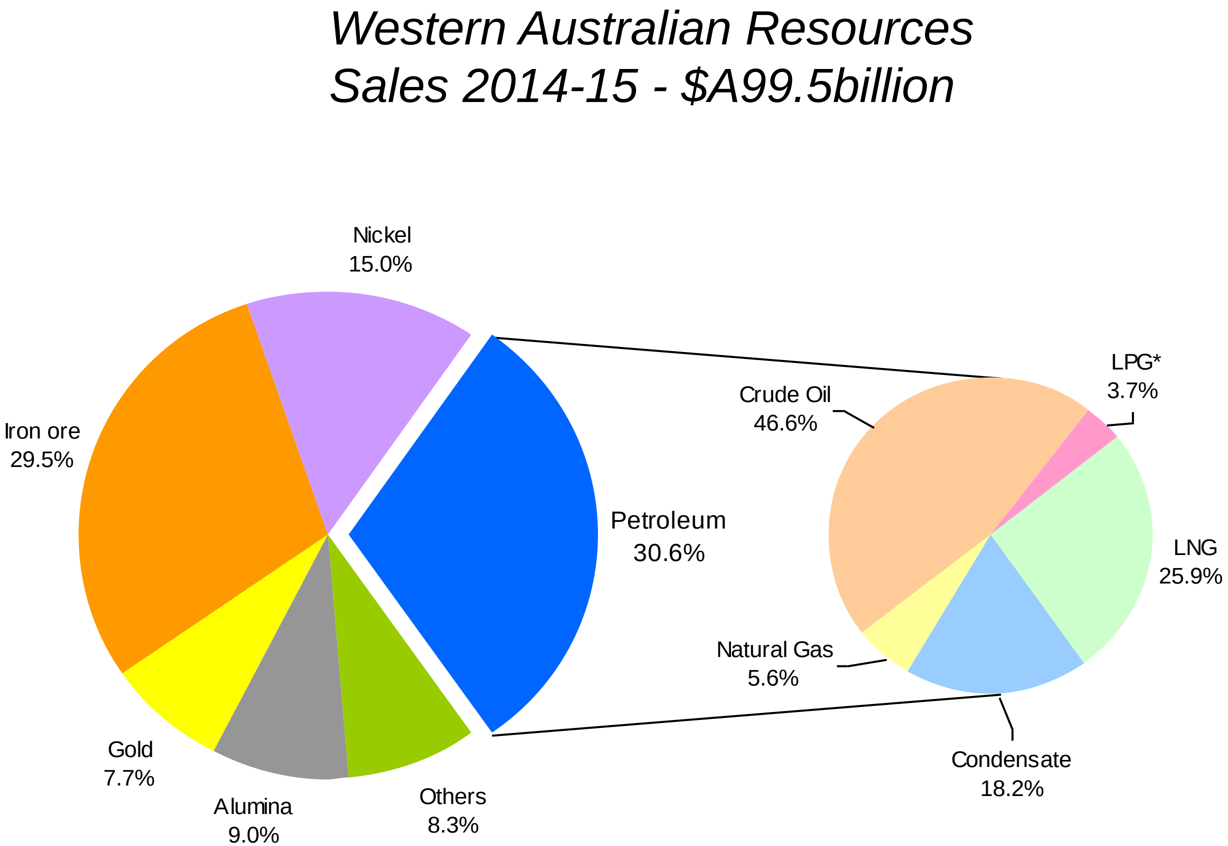
Example 2

We might wish to compare various combinations of drying times and amount of aggregate as to their effect of the strength of concrete. Certain test samples will have to be formed and their strengths measured. These measurements are then used in the inferential process of determining which combination of drying time and amount of aggregate produces the best product.

Example 3

A quality control engineer might periodically sample a few manufactured items coming off an assembly line and count the number of defectives. This number is then used to decide whether or not the line is operating within nominal standards.

Resource Sector of Western Australia



Source: Western Australian Department of Mines and Petroleum
http://www.dmp.wa.gov.au/

Talking Point

Newspaper reports:

Crime rate jumps 50% in your city

What do you infer?
What if the crime rate in your city was 4%?

Simpson's Paradox

Data on Survival of patients after surgery in two hospitals A and B

	Hospital A	Hospital B
Died	63(3%)	16 (2%)
Survive	2037	784
	2100	800

Which hospital seems to have a lower death rate?

Let's look at Patients Condition before surgery :

Survival data - good condition Patients

	Hospital A	Hospital B
Died	6(1%)	8 (1.3%)
Survived	594	592
	600	600

Survival data - poor condition Patients

	Hospital A	Hospital B
Died	57 (3.8%)	8 (4%)
Survived	1443	192
	1500	200

Which hospital seems to have a lower death rate?

Why this paradox?

Key Statistical Issues and Questions in an Investigation

- How was the data collected?
- How do we analyse it?
- How do we select the correct test?
- What information does it give us?
- What conclusions can we draw?

We will learn to answer these questions in the weeks that follow.

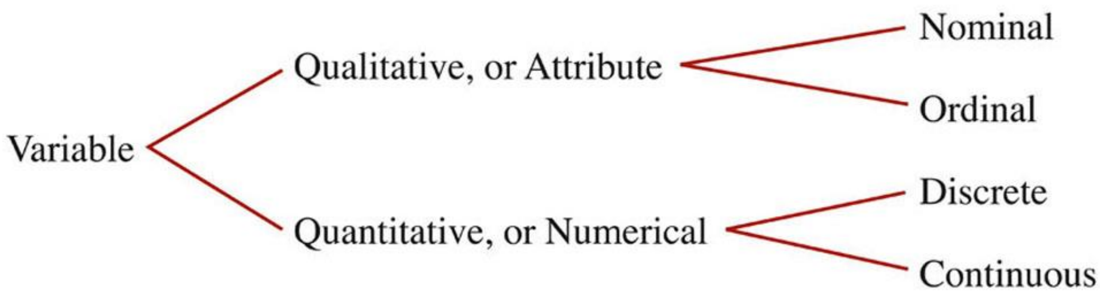
Some Terminology

- **Population**: The entire collection of observations of interest corresponding to individuals or objects under study.
- **Sample**: a subset of a population.
- **Statistic**: a numerical measure that describes an aspect of a sample.
- **Variable**: a characteristic of interest about each individual element of a population or sample to be measured or observed, e.g. height, weight.

Processing Data collected

- Raw data are not very informative.
- One of the aims of statistics is to obtain meaningful information from data.
- This can be done using either a numerical or graphical summary.
- Summary consistent with objective and limited by data/collection strategy.

Data Types



•**Qualitative or categorical data** is data that can be classified according to some attribute or characteristic (pass/fail, hair colour, etc)

•**Quantitative data** is data that is measured or counted. Operations such as addition and multiplication can be performed on quantitative data and give meaningful results.

•Quantitative data may be

- •**discrete** (count data)
- •**continuous** (measured)

Data type	Data type	Definition	Examples
Qualitative (Categorical)	Nominal	Categorised by names only	Colour, gender, species
Qualitative (Categorical)	Ordinal	Arranged in classes which themselves form an ordered sequence	Degree class
Quantitative	Interval	Individual data are compared to one another without the need to refer to membership of classes. One value may be subtracted from another to yield a sensible answer, but the origin of the scale is arbitrary, so they are not absolute quantities	Temperature in degrees Centigrade
Quantitative	Ratio	Referenced to a zero value, so that the two values retain the same ratio irrespective of the units in which they have been measured	Length, volume

Examples: Identify the type of data

1. Taos, Acoma, Zuni, and Cochiti are the names of four Native American pueblos from the population of names of all Native American pueblos in Arizona and New Mexico.

2. In a high school graduating class of 319 students, Jim ranked 25^{th}, Ian ranked 19^{th}, Roy ranked 10^{th}, and Julia ranked 4^{th}, where 1 is the highest rank.

3. Body temperatures (in degrees Celsius) of Black Bream in the Swan River.

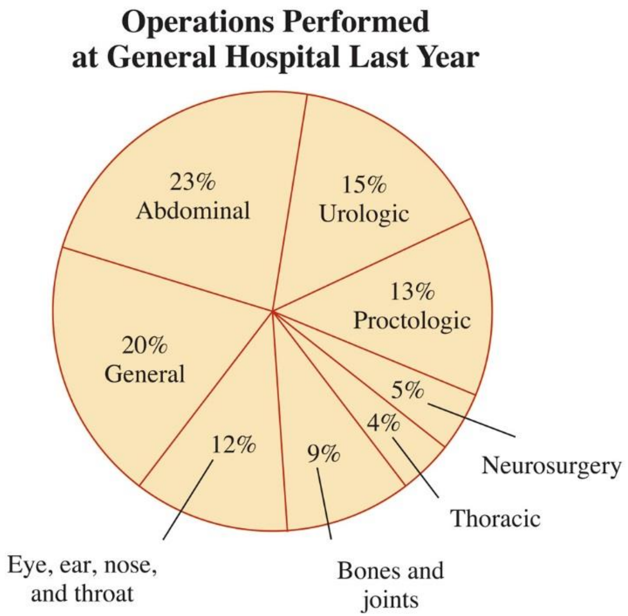
4. Length of Black Bream swimming in the Swan River.

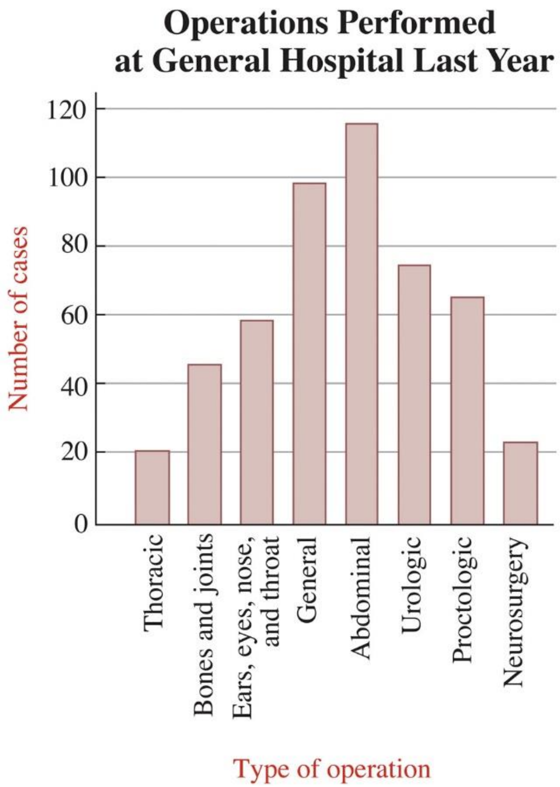
5. Data collected on the type of engineer (1 for electrical, 2 for chemical, 3 for mechanical, 4 for civil/industrial, and 5 for others)

Presentation of Quantitative data

Operations Performed at
General Hospital Last Year

Type of Operation	Number of Cases
Thoracic	20
Bones and joints	45
Eye, ear, nose, and throat	58
General	98
Abdominal	115
Urologic	74
Proctologic	65
Neurosurgery	23





Stem-and-Leaf Display

- A stem-and-leaf display (or stemplot) is a method of exploratory data analysis that is used to rank-order and arrange data into groups.
- It is useful for small amounts of data.
- It summarises and preserves the data at the same time.

How to make a stem-and-leaf display?

- Divide the digits of each data value into two parts. The leftmost part is called the *stem* and the rightmost part is called the *leaf*.
- Align all the stems in a vertical column from smallest to largest. Draw a vertical line to the right of all the stems.
- Place all leaves having the same stem in the same row as the stem, and arrange the leaves in increasing order.

Stem-and-Leaf Example

Sample of 19 Exam Grades

76	74	82	96	66	76	78	72	52	68
86	84	62	76	78	92	82	74	88	

Stem-and-Leaf display

5	2
6	2 6 8
7	2 4 4 6 6 6 8 8
8	2 2 4 6 8
9	2 6

Frequency Distributions and Histograms

Statistics Exam Scores

60	47	82	95	88	72	67	66	68	98
90	77	86	58	64	95	74	72	88	74
77	39	90	63	68	97	70	64	70	70
58	78	89	44	55	85	82	83	72	77
72	86	50	94	92	80	91	75	76	78

Range = highest score - lowest score
= 98-39=59

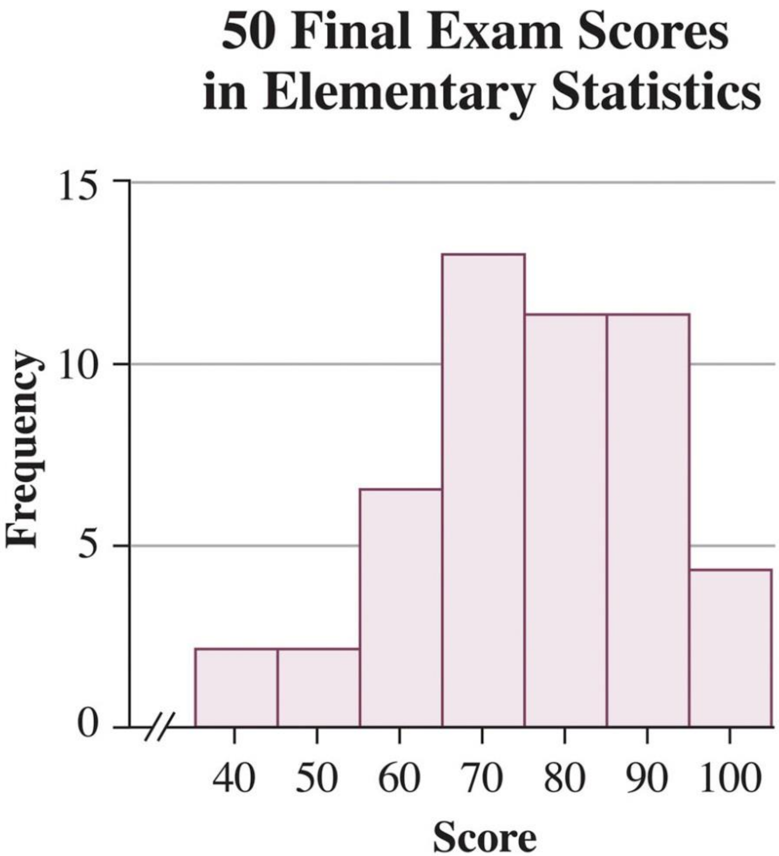
Number of classes = 7

Class width = 10

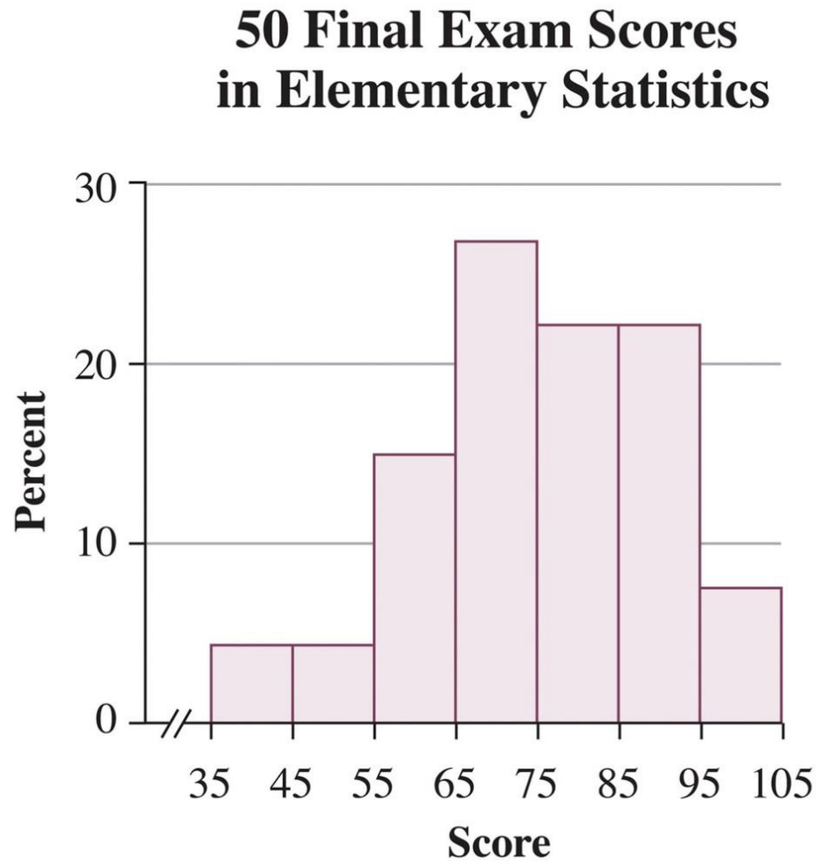
Frequency Distribution with Class Midpoints

Class Number	Class Boundaries	Frequency f	Class Midpoints x
1	$35 \leq x < 45$	2	40
2	$45 \leq x < 55$	2	50
3	$55 \leq x < 65$	7	60
4	$65 \leq x < 75$	13	70
5	$75 \leq x < 85$	11	80
6	$85 \leq x < 95$	11	90
7	$95 \leq x < 105$	4	100
		50	

Frequency histogram



Relative frequency histogram

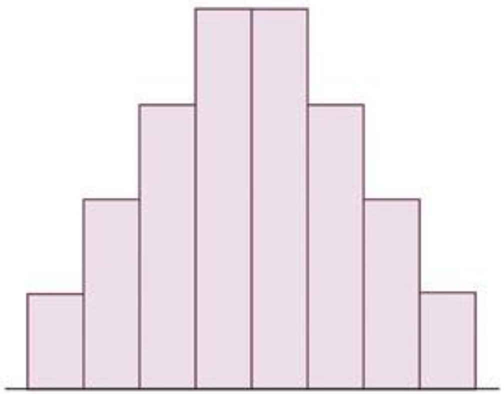


$$\text{Relative frequency} = \frac{\text{Class frequency}}{\text{Total of all frequencies}}$$

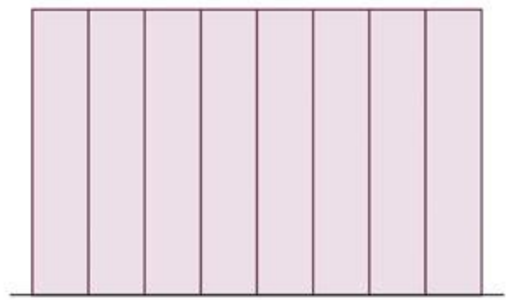
How to Construct a Histogram

1. Determine the sample size *n.*
2. Define *k* class intervals of equal width (usually 5 to 15, no universal rules for this).
3. Determine the frequency, f_i, of each class *i*.
4. Calculate the relative frequency (proportion) of each class: f_i/n.
5. For each class draw a bar whose width extends between corresponding class boundaries. The height of each bar corresponds to class frequency (or relative frequency).

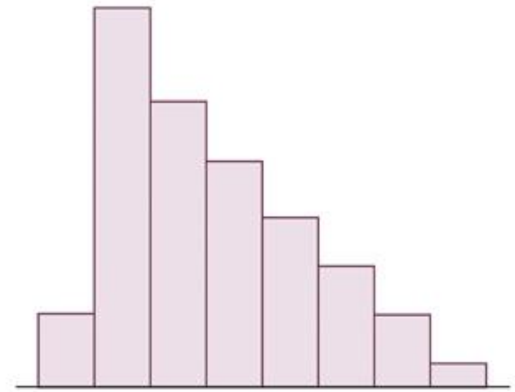
Shapes of Histograms



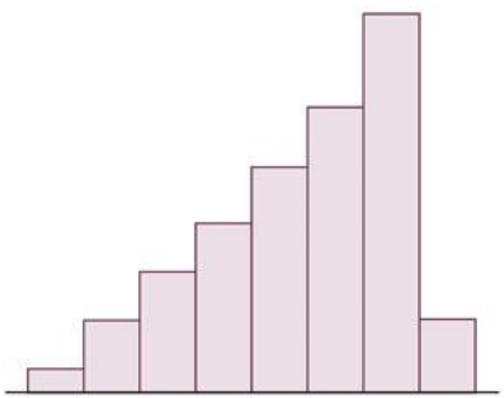
Symmetrical, normal, or triangular



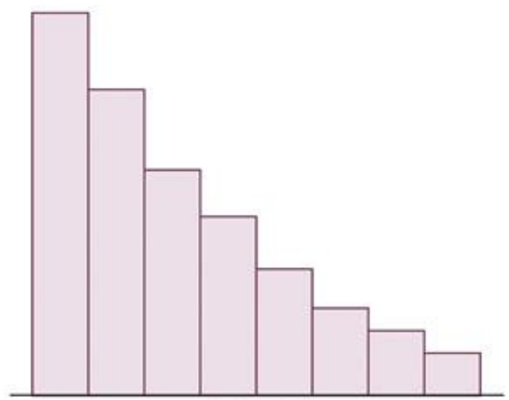
Symmetrical, uniform, or rectangular



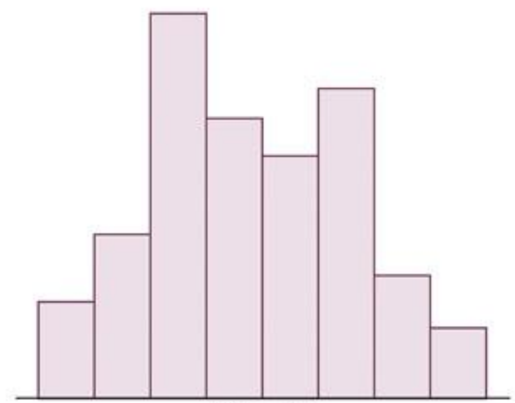
Skewed to right



Skewed to left



J-shaped



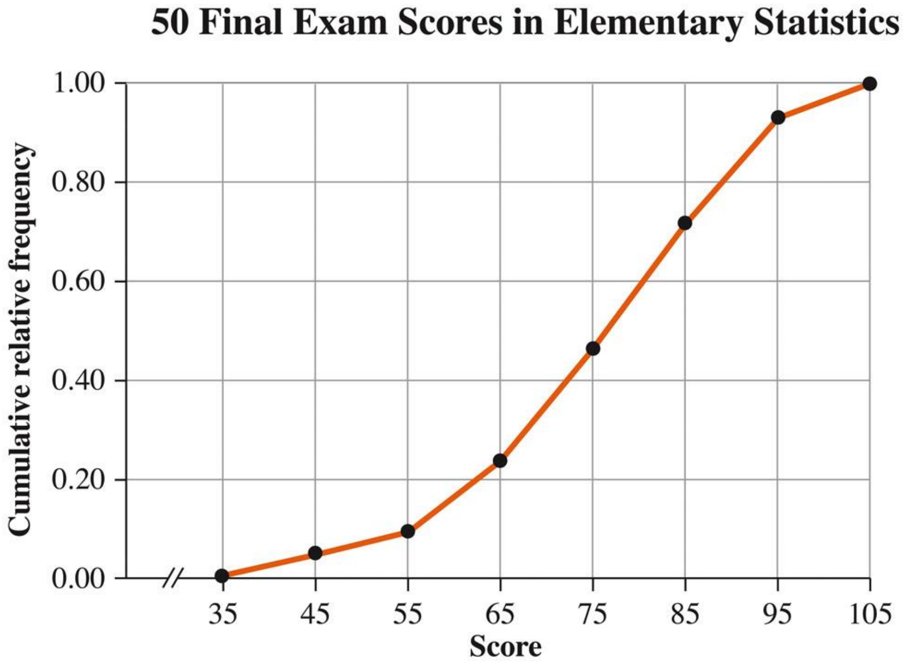
Bimodal

Cumulative Frequency Distribution and Ogives

Using Frequency Distribution to Form a Cumulative Frequency Distribution

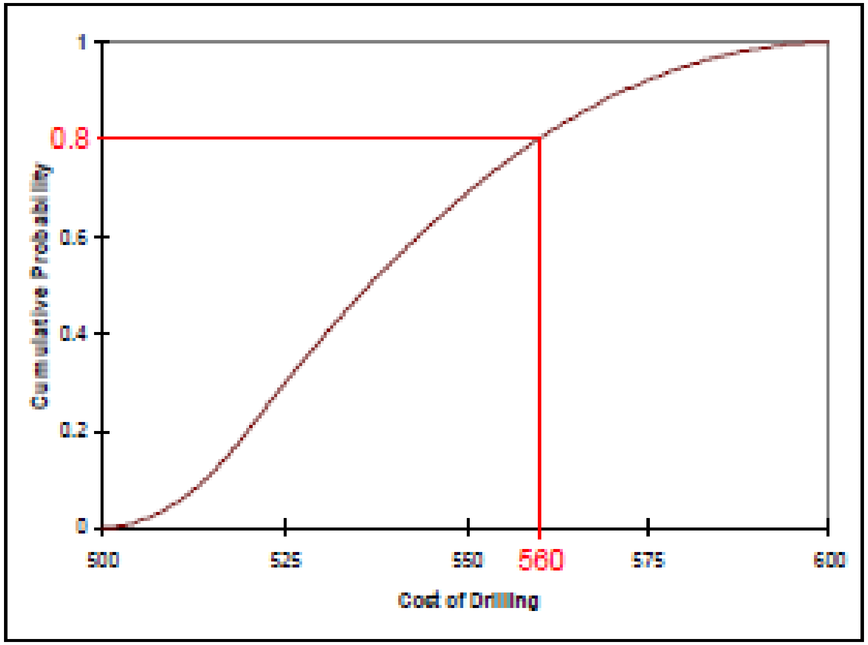
Class Number	Class Boundaries	Frequency f	Cumulative Frequency
1	$35 \leq x < 45$	2	2 (2)
2	$45 \leq x < 55$	2	4 (2 + 2)
3	$55 \leq x < 65$	7	11 (7 + 4)
4	$65 \leq x < 75$	13	24 (13 + 11)
5	$75 \leq x < 85$	11	35 (11 + 24)
6	$85 \leq x < 95$	11	46 (11 + 35)
7	$95 \leq x < 105$	4	50 (4 + 46)

Ogive



Cumulative Distribution

What are these curves used for?



Probability of realisation of values less than or equal to a given value

What is the probability that the cost of drilling would be less than or equal to 560?

Measures of Central Tendency

Values that locate, in some sense, the centre of a set of data.

- Mean
- Median
- Mode

Finding the Sample Mean

- **<u>Given</u>**: a set of data consisting of the five values

$$6, 3, 8, 6, 4$$

- Find the sample mean $\bar{x}$

The Formula – knowing its parts

- The calculation of a sample statistic requires the use of a formula. In this case, use:

$$\bar{x} = \frac{\sum x_i}{n}$$

- $\bar{x}$ is "*x-bar*", the sample mean
- $\sum x_i$ is the *"sum of x"*, the sum of all data
- n is the *"sample size"*, the number of data

Solution:

Sample = { 6, 3, 8, 6, 4 } ⟹ $n = 5$

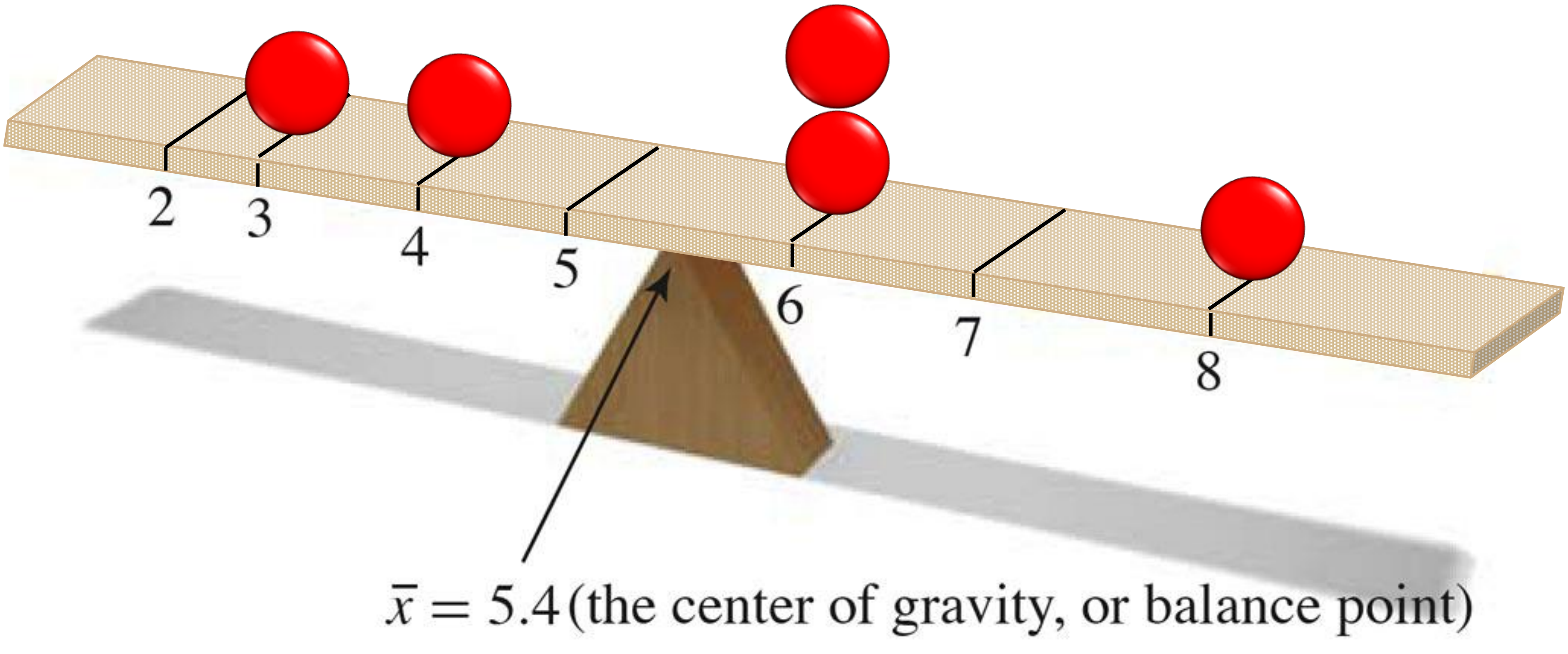
1 2 3 4 5

Hence,

$$\bar{x} = \frac{6+3+8+6+4}{5} = \frac{27}{5} = 5.4$$

<u>Therefore, the mean is 5.4.</u>

Physical Representation of the Mean



Finding the Sample Median

1. Order the data from smallest to largest
2. For an odd number of data values in the distribution,

$$\text{Median} = \text{Middle data value}$$

3. For an even number of data values in the distribution,

$$\text{Median} = \frac{\text{Sum of middle two values}}{2}$$

Example

- **Given**: The distance, in feet, ran in five seconds by preschoolers during a fitness evaluation test was recorded as:

6, 10, 13, 11, 12, 8, 8, 11

- Find the sample median.

Ranking the data

- Since the median is the “middle value”, the data must first be ranked in order of value
- Typically, ranking is smallest value first and largest value last:

Smallest 4th Largest 5th 7th 2nd 3rd 6th

Sample data = {~~6~~, ~~10~~, ~~13~~, ~~11~~, ~~12~~, ~~8~~, ~~8~~, ~~11~~}

Ranked data = { 6, 8, 8, 10, 11, 11, 12, 13}

1st 2nd 3rd 4th 5th 6th 7th 8th

Position of the middle value

- The position of the middle value (or depth) of the median is determined using the formula

$$\text{Position of middle value} = \frac{n+1}{2}$$

n = 8 $\Rightarrow$ *(8+1)/2 = 4.5*

So the two middle values are in the 4th and 5th positions of the ranked data.

Determining the Median Value

From the smallest value data →

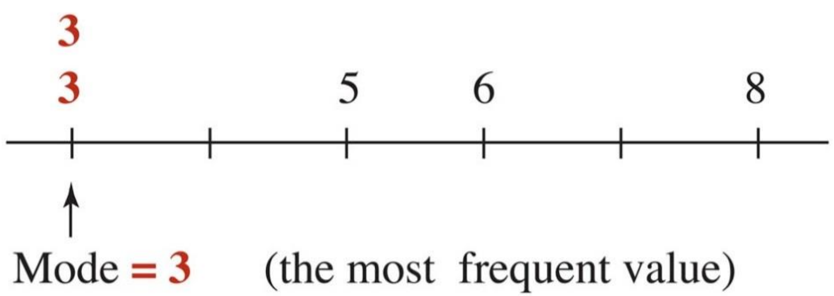
	Position 1	Position 2	Position 3	Position 4	Position 5			
Ranked data =	{6,	8,	8,	10,	11,	11,	12,	13}

$$\therefore \text{ Median} = \frac{10+11}{2} = \frac{21}{2} = 10.5 \text{ feet}$$

Finding the Mode

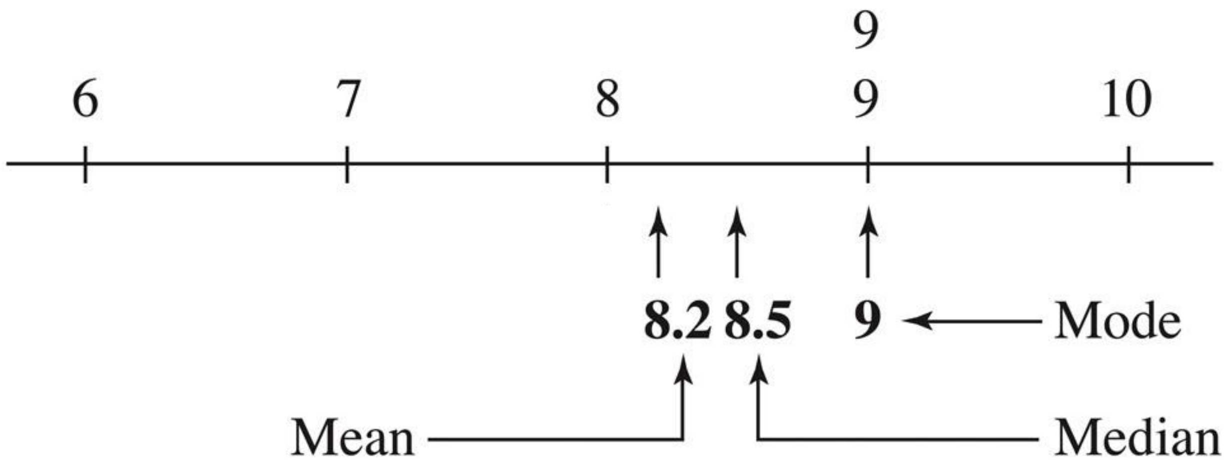
- The mode is the value of *x* that occurs most frequently

- Example: find the mode of {3,3,5,6,8}



Example

Measures of Central Tendency for {6,7,8,9,9,10)



Talking Point

Mean or Median

Country/Territory	Net mean wealth per adult	Gross mean wealth per adult	Net median wealth per adult	Debt per adult
Australia	402,578	503,070	219,505	100,492
Canada	251,034	313,186	90,252	62,151
Italy	241,383	266,731	138,653	25,348
Japan	216,694	251,733	110,294	35,039
New Zealand	182,548	231,867	76,607	49,319
United Kingdom	243,570	293,114	111,524	49,545
United States	301,140	357,951	44,911	56,811

http://www.international-adviser.com/ia/media/Media/Credit-Suisse-Global-Wealth-Databook-2013.pdf

Measures of Dispersion

- A measure of spread helps to tell us more about the data
- It gives a feel for how much variation there is in the data
- It tells us whether the values are clustered close to the mean (or median) or spread out.

Sample Standard Deviation

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

$$s = \sqrt{s^2}$$

- *s* is the sample standard deviation
- What is s^2 called?
- Why do we divide by n-1?
- Why do we take the square root?

Example

- **<u>Given</u>:** The times, in seconds, required for a sample of students to perform a required task were:

$$6, \quad 10, \quad 13, \quad 11, \quad 12, \quad 8$$

- **<u>Find</u>:** **a)** The sample variance, s^2

  b) The sample standard deviation, s

Solution

- Sample size n = 6
- The mean $\bar{x} = 10$

(a) Sample Variance is

$$s^2 = \frac{\sum(x_i - \bar{x})^2}{n-1} = \frac{(6-10)^2 + (10-10)^2 + (13-10)^2 + (11-10)^2 + (12-10)^2 + (8-10)^2}{6-1}$$

$$= \frac{34}{5} = 6.8$$

(Variance has no unit of measure, it's a number only)

(b) Standard deviation is

$$s = \sqrt{s^2} = \sqrt{6.8} = 2.60768 = 2.6 \text{ seconds}$$

Resistant Measures

A **Resistant Measure** is one that is not affected by outliers or highly skewed data.

When giving a numerical summary of the data we should give at least two values: a measure of centre and a measure of spread.

- The Median and IQR are resistant measures. Use these two when the data is skewed.
- The Sample Mean and the Sample Standard deviation are not resistant measures. Use these two for approximately symmetrical data.

Quartiles

Given a sample of n observations: $x_1, x_2, \ldots, x_n$, we can order them from smallest to largest resulting in the **order statistics**: $y_1 \leq y_2 \leq y_3 \leq \cdots \leq y_n$.

If $0 < p < 1$, then the **($100p$)th sample percentile** has approximately np observations less than it, and $n(1-p)$ sample observations greater than it.

- The 25^{th} percentile is called the **lower quartile** (Q_1). One quarter of observations lie below Q_1.
- The 50^{th} percentile is the median (Q_2)
- The 75^{th} percentile is called the **upper quartile** (Q_3). One quarter of observations lie above Q_3.
- 50% of observations lie between Q_1 and Q_3.

Finding the Lower and Upper Quartiles

The **lower** and **upper quartiles** are simply the (100*p*)th percentiles when $p = \frac{1}{4}$ and $p = \frac{3}{4}$, respectively.

- If $(n+1)p$ is an integer, then the $(100p)$th sample percentile is the $(n+1)p$ th order statistic.

- If $(n+1)p$ is not an integer, i.e. equals to $r + \frac{a}{b}$, then the $(100p)$th sample percentile is defined as: $y_r + \frac{a}{b}(y_{r+1} - y_r)$.

Example

The number of components per hour turned out on a lathe was measured on 14 occasions:

18	17	21	18	19	17	17
16	20	17	15	20	16	18

Find the lower and upper quartiles for the above sample.

Solution

First, we need to order the data:

15	16	16	17	17	17	17
18	18	18	19	20	20	21

- $n = 14$

- For Lower Quartile: $p = \frac{1}{4}$, $(n+1)p = \frac{15}{4} = 3\frac{3}{4}$, so $r = 3$, and $\frac{a}{b} = \frac{3}{4} \Rightarrow$ Lower quartile = $y_3 + \frac{3}{4}(y_4 - y_3) = 16 + \frac{3}{4}(17 - 16) = 16\frac{3}{4}$ =16.75

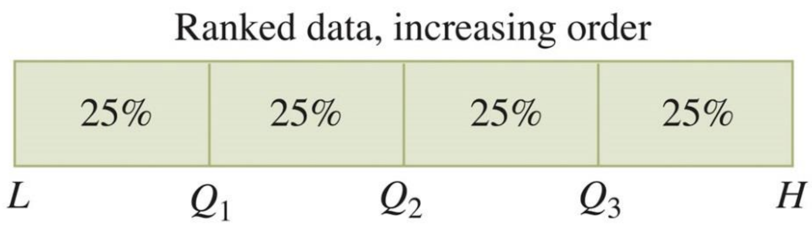
- For Upper Quartile: $p = \frac{3}{4}$, $(n+1)p = \frac{45}{4} = 11\frac{1}{4}$, so $r = 11$, and $\frac{a}{b} = \frac{1}{4} \Rightarrow$ Upper quartile = $y_{11} + \frac{1}{4}(y_{12} - y_{11}) = 19 + \frac{1}{4}(20 - 19) = 19\frac{1}{4}$ = 19.25

Range & Interquartile Range (IQR)

- Range = Max value – Min Value

- IQR = Q_3 - Q_1 .

- IQR gives a feel for how the middle 50% of the data is spread out.



5-Number Summary

- A 5-number summary consists of the numbers

  Minimum, Q_1, Median, Q_3, Maximum

- The five number summary for the set of 14 measurements in the previous example is:

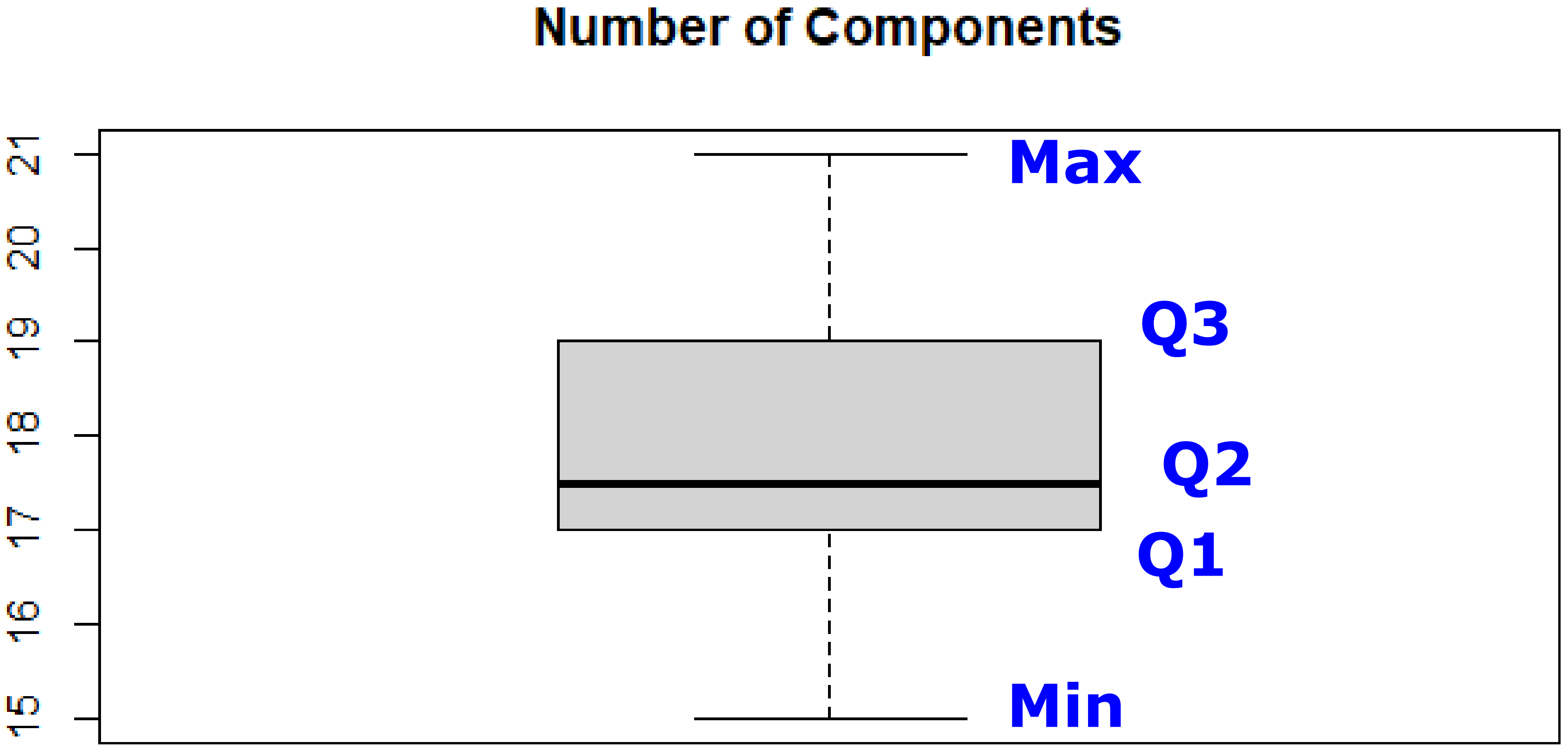
Min	Q_1	Median	Q_3	Max
15	16.75	17.5	19.25	21

Boxplot

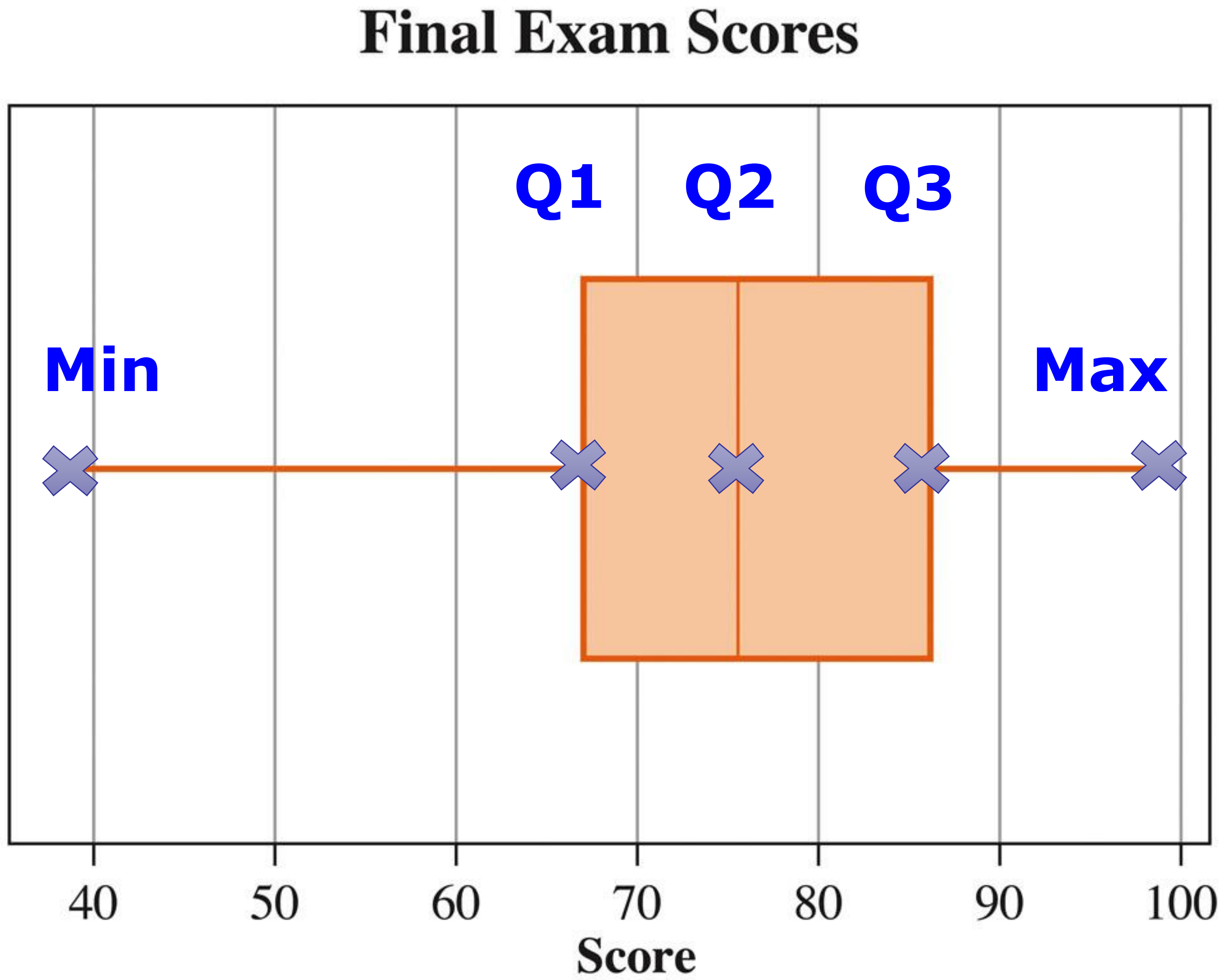
- In a boxplot Q_1 and Q_3 are the ends of the box. The vertical line in the box is the second quartile which is the median.

- The horizontal lines extending from the box are called whiskers. The whiskers will extend to the minimum and maximum values provided there are no outliers.

Boxplot



Boxplot (Example from Slides 22-23)



What is an Outlier?

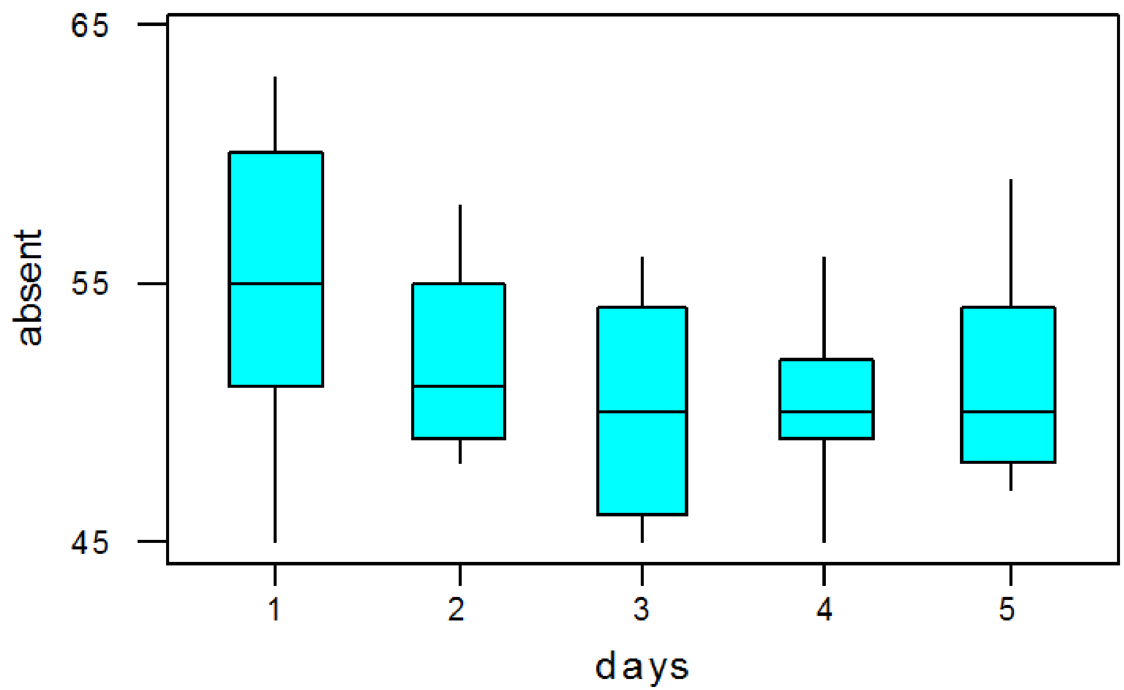
- An outlier is a value that does not fit in with the majority of the observations.
- It may be a perfectly valid observation or may have occurred because of some error in data collection or entry.
- You should investigate any outliers.

Outliers in Boxplots

- In a boxplot any observation more than 1.5(IQR) beyond the end of the box is recorded as an outlier, i.e. values less than $Q_1 - 1.5(IQR)$ and values more than $Q_3 + 1.5(IQR)$ are outliers.
- An outlier is marked with a star.
- The whiskers extend to the largest (smallest) value that lies within 1.5(IQR) from the ends of the box.

Boxplots for comparison of different samples

Side-by-side boxplots of absenteeism by day of week



Because each boxplot is so simple, they work well in side-by-side comparison across multiple samples.

Lecture Summary

Descriptive Statistics

- Sample mean $$\bar{x} = \frac{\sum x_i}{n}$$
- Sample Variance $$s^2 = \frac{\sum_i (x_i - \bar{x})^2}{n-1}$$
- Mode is the value that occurs most frequently
- Median is the middle value
- 5-number summary:

  Min Q_1 Median Q_3 Max

Data Presentation

- Step-and-leaf display
- Histogram
- Box plot