Chapter 4

Summarising Data

Recap and Outline

- Graphical methods of presenting data
- Numerical methods for summarising data
- Basic calculations
- MINITAB

Definitions

Algebraic Notation

1st random sample	1	5	7
2nd random sample	2	0	3
typical random sample	x_1	x_2	x_3

$$\sum_{i=1}^{n} x_i = x_1 + x_2 + \cdots + x_n$$

Definitions

Raising to powers:

$$x^k$$

Ordering with brackets: $\times \quad \div$ then $+ \quad -$

$$3 + 4^2 = 19$$

$$3^2 + 4^2 = 25$$

$$(3 + 4)^2 = 49.$$

In general

$$\sum x^2 \neq \left(\sum x\right)^2$$

Measures of Location

- The Mean
- The Median
- The Mode

The Mean ($\overline{x}$)

$$\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i \quad \text{or} \quad \frac{\sum x}{n}$$

The Mean ($\overline{x}$)

Date	Cars Sold	Date	Cars Sold
01/07/04	9	08/07/04	10
02/07/04	8	09/07/04	5
03/07/04	6	10/07/04	8
04/07/04	7	11/07/04	4
05/07/04	7	12/07/04	6
06/07/04	10	13/07/04	8
07/07/04	11	14/07/04	9

The mean number of cars sold per day is

$$\bar{x} = \frac{9+8+\ldots+8+9}{14} = 7.71.$$

The Mean ($\overline{x}$)

Cars Sold $(x_{(j)})$	**Frequency** (f_j)
4	1
5	1
6	2
7	2
8	3
9	2
10	2
11	1
Total (n)	14

The sample mean is

$$\overline{x} = \frac{4\times 1 + 5\times 1 + 6\times 2 + \ldots + 11\times 1}{14} = 7.71.$$

In general

$$\overline{x} = \frac{1}{n}\sum_{j=1}^{k} f_j x_{(j)}$$

The Mean ($\overline{x}$)

Data: sample mean is 9.73

8.4	8.7	9.0	9.0	9.2	9.3	9.3	9.5	9.6	9.6
9.6	9.7	9.7	9.9	10.3	10.4	10.5	10.7	10.8	11.4

Class Interval	**Mid Point** (m_j)	**Frequency** (f_j)
$8.0 \leq x < 8.5$	8.25	1
$8.5 \leq x < 9.0$	8.75	1
$9.0 \leq x < 9.5$	9.25	5
$9.5 \leq x < 10.0$	9.75	7
$10.0 \leq x < 10.5$	10.25	2
$10.5 \leq x < 11.0$	10.75	3
$11.0 \leq x < 11.5$	11.25	1
Total (n)		20

The Median

- Simply the "middle" observation (ordered)
- Odd number of observations (n):

$$\text{median} = \left(\frac{n+1}{2}\right)^{th} \text{ largest observation}$$

- Even number of observations (n):

$$\text{median} = \text{average of the } \left(\frac{n}{2}\right)^{th} \text{ and the } \left(\frac{n}{2}+1\right)^{th} \text{ largest observations}$$

The Mean ($\overline{x}$)

Can approximate the sample mean using

$$\bar{x} = \frac{1}{n}\sum_{j=1}^{k} f_j m_j.$$

For these grouped data

$$\bar{x} = \frac{1}{20}(1 \times 8.25 + 1 \times 8.75 + \cdots + 3 \times 10.75 + 1 \times 11.25)$$
$$= 9.775.$$

Close to correct value 9.73

Data:

8.4	8.7	9.0	9.0	9.2	9.3	9.3	9.5	9.6	9.6
9.6	9.7	9.7	9.9	10.3	10.4	10.5	10.7	10.8	

Sample size $n = 19$ is **odd**

$$\text{median} = \left(\frac{n+1}{2}\right)^{th} \text{ largest observation}$$
$$= 10^{th} \text{ largest observation}$$
$$= 9.6$$

Data:

8.4	8.7	9.0	9.0	9.2	9.3	9.3	9.5	9.6	9.6
9.6	9.7	9.7	9.9	10.3	10.4	10.5	10.7	10.8	11.4

Sample size $n = 20$ is **even**

$$\begin{aligned}\text{median} &= \text{average of the } \left(\frac{n}{2}\right)^{th} \text{ and} \\ &\quad \text{the } \left(\frac{n}{2}+1\right)^{th} \text{ largest observations} \\ &= \text{average of the } 10^{th} \text{ and the } 11^{th} \text{ largest observations} \\ &= \frac{9.6+9.6}{2} \\ &= 9.6\end{aligned}$$

The Median

- Possible to estimate from an ogive
- The median is the x-value corresponding to 50% cumulative frequency

The Mode

- Discrete data: the most common value
- Continuous data: the most common class

Class	Frequency
$10 \le x < 20$	10
$20 \le x < 30$	15
$30 \le x < 40$	30

Modal class is $30 \le x < 40$

Measures of Spread

- Location is not sufficient
- Need some idea of the spread of the data

The Range

- The difference between the largest and smallest values

$$Range = max - min$$

- Not the best measure of spread

The Quartiles

Lower quartile

$$Q1 = \frac{(n+1)}{4}\text{th smallest observation}$$

Upper quartile

$$Q3 = \frac{3(n+1)}{4}\text{th smallest observation}$$

The Inter-Quartile Range

- The range of the middle half of the data.

- Divide data into four sections separated by *quartiles*

  – *Lower quartile*, Q1 has 25% of the data below it

  – *Median*, Q2 has 50% of the data below it

  – *Upper quartile*, Q3 has 75% of the data below it

Data: $n = 20$

8.4	8.7	9.0	9.0	9.2	9.3	9.3	9.5	9.6	9.6
9.6	9.7	9.7	9.9	10.3	10.4	10.5	10.7	10.8	11.4

Lower quartile

$$\begin{aligned} Q1 &= \frac{(n+1)}{4}\text{th smallest observation} \\ &= 5\frac{1}{4}\text{th smallest observation} \\ &= 9.225 \end{aligned}$$

Upper quartile

$$\begin{aligned} Q3 &= \frac{3(n+1)}{4}\text{th smallest observation} \\ &= 15\frac{3}{4}\text{th smallest observation} \\ &= 10.375 \end{aligned}$$

The Inter-Quartile Range

The Inter-Quartile Range is the difference between the upper and lower quartiles:

$$IQR = Q3 - Q1$$

Can approximate the sample variance from grouped frequency data using

$$s^2 = \frac{1}{n-1}\left\{\sum_{i=1}^{k} f_i m_i^2 - n\,(\bar{x})^2\right\}$$

The Sample Standard Deviation (s)

$$\begin{aligned} Standard\ Deviation &= \sqrt{Variance} \\ s &= \sqrt{s^2} \end{aligned}$$

Calculator: use σ_{n-1} or s buttons NOT σ_n or σ buttons

The Sample Variance (s^2)

The average of the squared distances of the observations from the mean:

$$s^2 = \frac{(x_1-\bar{x})^2 + (x_2-\bar{x})^2 + \ldots + (x_n-\bar{x})^2}{n-1}$$

General formula

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

or equivalently

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

Data: $n = 20$, sample mean is $\bar{x} = 9.73$

8.4	8.7	9.0	9.0	9.2	9.3	9.3	9.5	9.6	9.6
9.6	9.7	9.7	9.9	10.3	10.4	10.5	10.7	10.8	11.4

$$\begin{aligned} \sum x^2 &= 8.4^2 + 8.7^2 + \cdots + 11.4^2 = 1904.38 \\ n(\bar{x})^2 &= 1893.458 \end{aligned}$$

Sample variance is

$$\begin{aligned} s^2 &= \frac{1}{n-1}\left\{\sum_{i=1}^{n} x_i^2 - n\,(\bar{x})^2\right\} \\ &= \frac{1}{19}(1904.38 - 1893.458) = 0.57484 \end{aligned}$$

Sample standard deviation is

$$s = \sqrt{s^2} = \sqrt{0.57484} = 0.75818.$$

Summary statistics in MINITAB

MINITAB can be used to calculate many of basic numerical summary statistics described so far using

`Stats > Basic Statistics > Display Descriptive Statistics`

Box and Whisker Plots

Plot of summary statistics from data:

- Minimum (min)
- Lower quartile ($Q1$)
- Median ($Q2$)
- Upper quartile ($Q3$)
- Maximum (max)