Chapter 1 Getting Started

Data

Data is a set of numeric observations.

Sources of data:

- Government statistical agencies (Statistics Canada)
- Stock market activity
- Surveys

Population and Sample

The **population** is the complete set of numeric observations.

A **sample** is an observed subset of observations taken from the population.

A **random sample** contains a set of observations that are representative of the population.

Economic data sets are viewed as a sample from the population.

The challenge is to use the sample to make statements about the behaviour of the population. Some degree of uncertainty must be recognized since the sample is only one snapshot of the population.

To get started with exploring and describing the data consider:

- Graphical presentation of data gives a visual display.
- Statistics give numerical summary measures of the information in the data.

Chapter 1.2 Variables

A **variable** is any characteristic of a population or sample that is of interest to study. Data are the numerical observations.

Example

A data set contains facts about a sample of 79 companies selected from the Forbes 500 list for the year 1986.
The variables in the data set are:

1. Amount of assets (in millions $)
2. Amount of sales (in millions $)
3. Number of employees (in thousands)
4. Sector code:
    1 Energy
    2 Transportation
    3 Communication
    4 HiTech
    5 Finance
    6 Retail
    7 Manufacturing
    8 Medical
    9 Other

The numeric observations for the first 20 companies in the data set are:

Assets	Sales	Employees	Sector code
2687	1870	18.2	9
13271	9115	143.8	9
13621	4848	23.4	1
3614	367	1.1	5
6425	6131	49.5	2
1022	1754	4.8	4
1093	1679	20.8	7
1529	1295	19.4	9
2788	271	2.1	5
19788	9084	79.4	3
327	542	2.8	5
1117	1038	3.8	1
5401	550	4.1	5
1128	1516	13.2	7
1633	701	2.8	1
44736	16197	48.5	5
5651	1254	6.2	1
5835	4053	10.8	1
278	205	3.8	8
5074	2557	21.9	3

Classification of Variables

- Quantitative variables

  Continuous – the numeric observations can take any value in some interval.

  Example: for the data set of large companies assets, sales and number of employees are all continuous variables.

  Discrete – the numeric observations can take a limited number of values.

  Example 1: for a survey of students in Economics 325 the student age in years is a discrete variable with typical values 18, 19, 20, 21 etc.

  Example 2: for a survey of small businesses (say less than 50 employees) the number of employees may be viewed as a discrete variable.
  Note ~ A variable that takes on enough discrete values may be viewed as continuous for practical work.

- Qualitative or categorical variables – responses do not have a numerical meaning.

  Example 1: for the data set of large companies the sector code is a categorical variable.

  Example 2: for a survey of students in Economics 325 categorical variables include gender (Male/Female), mode of transportation to UBC (bus, car, bicycle, walk), etc.

Units of measurement for quantitative variables

- Levels

  Example 1: for the company data set assets is reported in millions of $ and number of employees is reported in thousands.

  Example 2: for stock market data, the closing daily price of a stock of a company is the price of one share of the stock quoted in $.

- Proportions

  Example: for a company stock denote the closing price on day t as $\mathbf{p_t}$. The daily return is defined as the proportionate change:

$$\frac{\mathbf{p_t - p_{t-1}}}{\mathbf{p_{t-1}}}$$

- Percentages

  Example: for a company stock the percentage daily return is defined as:

$$\mathbf{100 \cdot \frac{p_t - p_{t-1}}{p_{t-1}}}$$

Chapter 2 Describing Data

For a variable, the population has **N** observations.

The sample has **n** observations denoted by:

$$x_1, x_2, \ldots, x_n$$

n is the **sample size** of the data set.

The sum of the sample observations is:

$$x_1 + x_2 + \cdots + x_n$$

Using summation notation the sum is written as:

$$\sum_{i=1}^{n} x_i$$

If the population is available, the sum of the observations for the population is:

$$\sum_{i=1}^{N} x_i \qquad \text{where } N > n$$

Chapter 2.1 Measures of Central Tendency

(1) Mean

The **population mean** is a summary measure called a parameter and is denoted by the Greek letter μ (mu). This is calculated as:

$$\mu = \frac{1}{N} \sum_{i=1}^{N} x_i$$

Typically, economic data sets are viewed as a sample (the population is unknown or not available).
The **sample mean** is called a statistic and is denoted by $\bar{x}$ (x-bar).
This is calculated as:

$$\bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i$$

(2) Median

The **sample median** is the middle value of the sample observations. The calculation is simple – sort the observations and pick the middle value so half the sample observations are greater than the median and half the observations are smaller (when the sample size n is even, take the average of the two middle values).

Mean versus median – when the sample data are distributed symmetrically about the central value the mean and median will be identical.
In contrast, for data sets with a few extremely large observations the mean will exceed the median. That is, the mean can be very influenced by extreme observations – the median is not so affected.

> Example 1: consider a survey of household incomes in Canada for the current year. The mean may be pushed up by a relatively small group of 'wealthy' households to give an overly optimistic picture of the economic well-being of a typical Canadian household. In this case, the median may be a useful summary measure to report.
>
> Example 2: for the 1986 data set on 79 companies described earlier sample statistics for sales were calculated as:
>
> mean 4,178 millions of $
>
> median 1,754 millions of $

(3) Mode

The **mode** is the numerical value that occurs most frequently.
This summary statistic has application for data sets with repeating observations such as discrete or categorical variables.

> Example : for the 1986 data set on 79 companies the sector code has a mode of 5 – the code for the finance sector. Inspection of the data shows that 17 companies were allocated to the finance sector. The other sectors were represented by fewer companies.

(4) Geometric mean

For data sets with positive observations the **geometric mean**, denoted by $\bar{x}_g$, is defined as:

$$\bar{x}_g = \left(\prod_{i=1}^{n} x_i \right)^{1/n} \qquad \text{where } \prod \text{ is the product operator}$$

For calculation, it can be noted:

$$\begin{aligned} \log(\bar{x}_g) &= \frac{1}{n} \log\left(\prod_{i=1}^{n} x_i \right) \\ &= \frac{1}{n} \sum_{i=1}^{n} \log(x_i) \end{aligned}$$

A result from calculus is: $\bar{x}_g \le \bar{x}$ (the arithmetic mean)

Example: Annual growth rates in the price of a company stock for the past 5 years are:

4.3%, 6.0%, 3.5%, 8.2%, 7.0%

A 4.3% growth rate means that for a price p_0 the next period price p_1 is $p_1 = (1+0.043)\, p_0 = 1.043\, p_0$. The number 1.043 is called an index number. For the five years of data the growth in each period is expressed by the index numbers:

1.043, 1.06, 1.035, 1.082, 1.07

The mean growth over the 5-year period is calculated as the geometric mean:

$$\bar{x}_g = (1.043 \cdot 1.06 \cdot 1.035 \cdot 1.082 \cdot 1.07)^{1/5} = \mathbf{1.05786}$$

The mean growth rate can be stated as:

$$\mathbf{100 \cdot (\bar{x}_g - 1) = 5.786\%}$$

Now suppose that the growth rate of 5.786% continued.
The number of years required for the price to double is the value $\mathbf{t}$ such that:

$$\mathbf{1.05786^t = 2}$$

To solve for $\mathbf{t}$ take logarithms of both sides to get:

$$\mathbf{t \cdot \log(1.05786) = \log(2)}$$

The solution is:

$$\mathbf{t = \frac{\log(2)}{\log(1.05786)} = 12.32}$$

Problem: Give an example of a data set where the mean, median, mode and geometric mean are all identical.

Chapter 2.2 Measures of Variability

Example: Two sections of a course are given by two different instructors. It is of interest to consider if the student performance is similar in the two sections.

Summary statistics for the final grades are:

	Mean	Minimum	Maximum
Instructor A	72	40	98
Instructor B	72	55	92

Both sections have the same mean.
But, by inspecting the minimum and maximum grades, Instructor A's course has greater spread in the grades to suggest more variability than Instructor B's course.

Now consider alternative measures of variability.

A measure of total spread in the data (variability or dispersion) is the **range** defined as:

Range = Maximum – Minimum

A widely applied measure of variability is the **variance**.
For a sample of data $x_1, x_2, \ldots, x_n$, look at the squared distances between each observation and the sample mean:

$$(x_1 - \bar{x})^2,\ (x_2 - \bar{x})^2, \ldots, (x_n - \bar{x})^2$$

where the sample mean is calculated as: $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$

The use of squared distances ensures that a negative distance (the observation is smaller than the sample mean) is treated exactly the same as a positive distance (the observation is greater than the sample mean).

The **sample variance** is defined as:

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

Note that the divisor is n–1. This is a rule for working with the sample variance.

If the sample size n is relatively large then:

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

where the $\cong$ operator means "approximately equal to".

The use of squared distances in the variance calculation means that the units of measurement of the data are changed.

To restore the data to the original units of measurement consider the **sample standard deviation** defined as:

$$s = \sqrt{s^2}$$ (the positive square root is used).

The **population variance** is a parameter denoted by σ^2 (sigma-squared). This is defined as:

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

where μ is the population mean.

The **population standard deviation** is: $\sigma = \sqrt{\sigma^2}$

An alternative calculation formula for the sample variance can be stated. Note that:

$$\begin{aligned}\sum_{i=1}^{n}(x_i - \bar{x})^2 &= \sum_{i=1}^{n}\left(x_i^2 - 2x_i\bar{x} + \bar{x}^2\right)\\ &= \sum_{i=1}^{n}x_i^2 - 2\bar{x}\sum_{i=1}^{n}x_i + n\bar{x}^2\\ &= \sum_{i=1}^{n}x_i^2 - 2n\bar{x}^2 + n\bar{x}^2\\ &= \sum_{i=1}^{n}x_i^2 - n\bar{x}^2\end{aligned}$$

The above used the result: $\sum_{i=1}^{n}x_i = n\bar{x}$

Therefore, a calculation formula for the variance is:

$$s^2 = \frac{1}{n-1}\left[\sum_{i=1}^{n}x_i^2 - n\bar{x}^2\right]$$

The variance can be used to compare the variability of two samples that come from populations with the same mean.
However, to compare samples that come from populations with different means an adjusted measure of variability is needed.

The **sample coefficient of variation** gives an adjusted measure of variability defined as:

$$\frac{s}{\bar{x}} = \frac{\text{sample standard deviation}}{\text{sample mean}}$$

An alternative reporting style for the coefficient of variation is to express the standard deviation as a percentage of the mean:

$$\frac{s}{\bar{x}} \cdot 100$$

The coefficient of variation is a meaningful statistic for data sets where all observations are positive.

Example : for the company survey data set for 1986 consider comparing sales for companies operating in the finance sector with sales for companies in the manufacturing sector.

Summary statistics are presented in the table:

	n	$\bar{x}$	s
Finance	17	\$ 1,781 millions	3805
Manufacturing	10	\$ 4,925 millions	4097

When comparing two data sets the standard deviation for the sample with the higher mean is typically higher.

The coefficient of variation reveals a different picture:

	$s / \bar{x}$
Finance	2.136
Manufacturing	0.832

The sample range and sample variance may be distorted by unusual extreme observations. To avoid this, discard some of the extreme observations and find the range of those remaining.
A way to proceed is as follows.

Sort the sample observations in ascending order and denote the ordered observations as:

$$x_{(1)}, x_{(2)}, \ldots, x_{(n)}$$
$$\uparrow \qquad\qquad \uparrow$$
minimum maximum

Define the sample statistics:

$Q_1 = x_{((n+1)/4)}$ first quartile or 25th percentile

$Q_2 = x_{((n+1)/2)}$ second quartile or median

$Q_3 = x_{(3(n+1)/4)}$ third quartile or 75th percentile

If (n+1) is not an integer multiple of 4 then some adjustment or interpolation between two neighbouring observations is needed.

For a sample of data, the **interquartile range** is defined as:

$$Q_3 - Q_1$$

This statistic gives the spread of the middle 50% of the observations. This measure of variability is very little influenced by an occasional extreme observation.

Example: For a sample of 9 observations the ordered data is:

$x_{(1)}$	$x_{(2)}$	$x_{(3)}$	$x_{(4)}$	$x_{(5)}$	$x_{(6)}$	$x_{(7)}$	$x_{(8)}$	$x_{(9)}$
7	10	13	15	16	24	25	28	30

The sample median is $x_{(5)} = 16$.

The first quartile lies between observations $x_{(2)}$ and $x_{(3)}$.
A solution is to take the average of these two observations to get:

$Q_1 = (10+13)/2 = 11.5.$

The third quartile lies between observations $x_{(7)}$ and $x_{(8)}$ and can be calculated as:

$Q_3 = (25+28)/2 = 26.5.$

This gives the interquartile range:

$$Q_3 - Q_1 = 26.5 - 11.5 = 15$$

Chapter 2.4 Measures of Relationships between Variables

Economic theory proposes relationships among two or more variables.

Consider a data set that is a random sample of n observations on two variables. The numeric observations are denoted by:

$$(x_i, y_i) \qquad \text{for} \quad i = 1, 2, \ldots, n$$

The sample means are: $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$ and $\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$

and the sample variances are s_x^2 and s_y^2.

The **sample covariance** gives a measure of a linear association between two variables (that is, a scatter plot of the observations follows a straight line – either upward sloping or downward sloping) and is defined as:

$$s_{xy} = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})$$

An alternative calculation formula can be stated. Note that:

$$\begin{aligned}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}) &= \sum_{i=1}^{n}(x_i y_i - x_i\bar{y} - y_i\bar{x} + \bar{x}\bar{y}) \\ &= \sum_{i=1}^{n} x_i y_i - n\bar{x}\bar{y} - n\bar{x}\bar{y} + n\bar{x}\bar{y} \\ &= \sum_{i=1}^{n} x_i y_i - n\bar{x}\bar{y}\end{aligned}$$

Therefore, $$s_{xy} = \frac{1}{n-1}\left[\sum_{i=1}^{n} x_i y_i - n\bar{x}\bar{y}\right]$$

The sample covariance depends on the units of measurement of the two variables. To get a scale-free measure of the strength of a linear association use the **sample correlation** defined as:

$$r = \frac{s_{xy}}{s_x s_y}$$

where s_x and s_y are the sample standard deviations.

A result is: $-1 \le r \le 1$ or $|r| \le 1$

The interpretation is that as $|r|$ gets closer to one the stronger the evidence for a linear relationship between the two variables.

- $r > 0$ indicates a positive linear relationship
- $r < 0$ indicates a negative linear relationship
- $r = 0$ indicates no linear relationship – the variables are uncorrelated
- $r = 1$ gives a perfect positive linear relationship – the observations exactly fit on an upward sloping straight line.

Example: For the company data set introduced earlier consider the relationship between sales (in millions $) and number of employees (in thousands) for companies with number of employees not exceeding 25,000. The number of companies in the sample is n=51. Calculations show:

Covariance = 4219.1 Correlation = 0.617

Now consider the effect of rescaling the observations.
For each company divide the sales by 1000 so that the data is now expressed in billions $. Calculations now show:

Covariance = 4.219 Correlation = 0.617

The rescaling changed the numeric value for the covariance but the correlation remained the same.

The relationship between sales and employees
A scatter plot of the observations