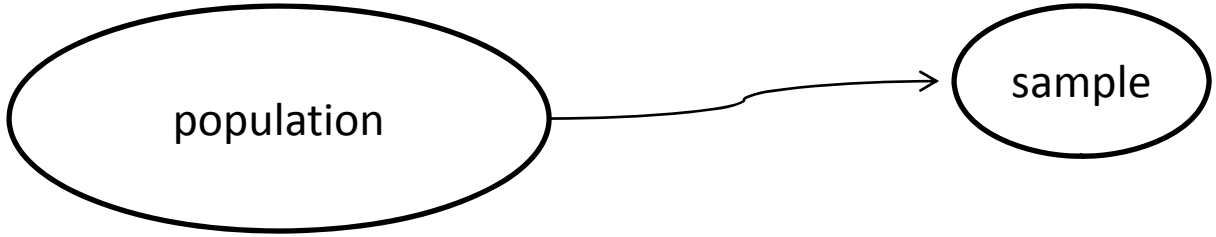
Introduction to Statistical Modelling

STAT2507A

Chapter 1

Describing Data with Graphs

POPULATION



- **Population**
  - Population is the set of all measurements of interest to the investigator
    - Examples: All Canadians, All students at Carleton, all students in STAT2507.
  - Interested in information in the population, but it might be
    - too expensive or too time consuming.
      - Example: measuring the body temperature of all Canadian
    - impossible to enumerate (enumerating the population would destroy it)
      - Example: measuring the volume of content in a pop can.

SAMPLE

- **Sample**
  - Sample is a subset of the measurements selected from the population of interest.
    - Examples: randomly selected 1000 Canadians, 100 students from Carleton.
  - Try to describe or predict the behaviour of the population on the basis of information obtained from representative sample from that population.

DESCRIPTIVE STATISTICS

- Set of measurements from population or sample needs to be summarized or organized.
  - Branch of statistics that presents techniques for describing sets of measurements is called **descriptive statistics**.
    - **Descriptive Statistics**: consists of procedures used to summarize and describe the important characteristics of a set of measurements.

INFERENTIAL STATISTICS

- When only sample from the population is available, need to answer question about population by using information from sample.
  - Branch of statistics that deals with this situation is called **inferential statistics**.
    - **Inferential Statistics:** consists of procedures used to make inferences about population characteristics from information contained in a sample drawn from this population
  - Objective of inferential statistics is to make inferences about the characteristics of a population from information contained in a sample.

VARIABLE

- **Variable:**
  - It is a characteristic that changes or varies over time and/or for different individuals or objects under consideration
  - Examples: Body temperature (varies over time for a certain individual) and it varies from person to person, hair colour, white blood cell count

EXPERIMENTAL UNIT

- **Experimental Unit:**
  - It is the individual or object on which a variable is measured.
  - A single measurement of data value results when a variable is actually measured on an experimental unit
  - A set of measurements, called **data**, can be either a **sample** or a **population**

EXAMPLES

- Example 1
  - Variable: Hair colour
  - Experimental unit: person
  - Typical measurements: black, blonde, brown etc.

- Example 2
  - Variable: Time until light bulb burns out
  - Experimental unit: bulb
  - Typical measurements: 1500 hours, 1535.5 hours etc.

EXERCISE

1.1) Identify the experimental units and type of measurements on which the following variables are measured:

a) Gender of a student
b) Number of errors on a midterm exam with 10 questions
c) Age of a cancer patient
d) Colour of a car entering the parking lot

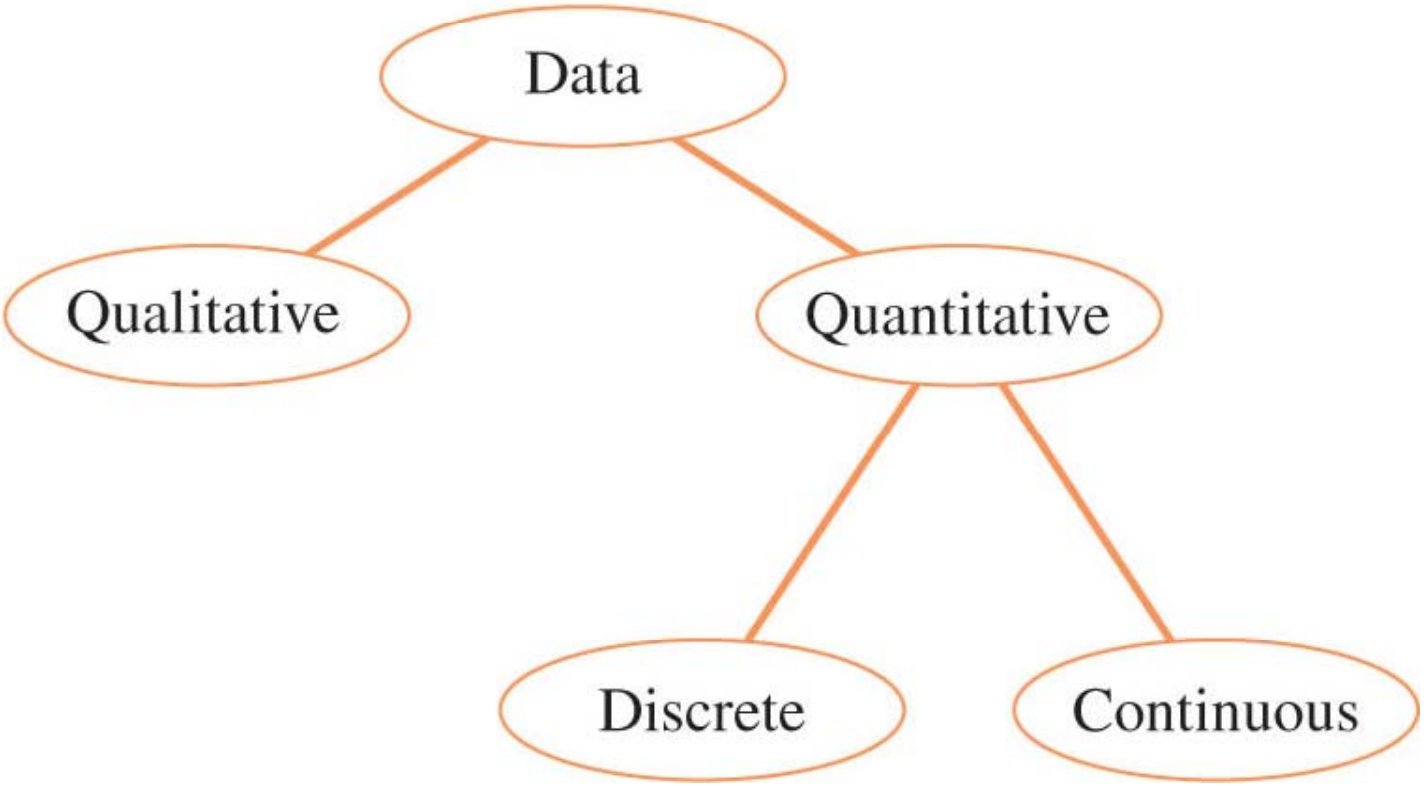
How many variables have you measured?

- Univariate data
  - Univariate data results when a single variable is measured on a single experimental unit
- Bivariate data
  - Bivariate data results when two variables are measured on a single experimental unit
- Multivariate data
  - Multivariate data results when more than two variables are measured on a single experimental unit.

EXAMPLES

- Univariate data: Measuring body temperature of each of 100 people
- Bivariate data: Measuring body temperature and white blood cell count of each of 100 people
- Multivariate data: Recording gender, height, measuring body temperature and white blood cell count of each of 100 people

TYPES OF VARIABLES



QUALITATIVE (CATEGORICAL) VARIABLES

- **Qualitative (categorical) variables** measure a quality or characteristics on each experimental unit.
  - Hair colour (black, brown, blonde…)
  - Make of car (Dodge, Honda, Ford…)
  - Gender (male, female)
  - Province of birth (Alberta, Ontario…)

QUANTITATIVE VARIABLES

- **Quantitative variables** measure a numerical quantity or amount on each experimental unit
  - **Discrete**: if it can assume only a finite or countable number of values.
  - **Continuous**: if it can assume infinitely many values corresponding to the points on a line interval.

EXAMPLES

- Qualitative Variables
  - Letter grade of each student in the class: A^+, A, A^-, B^+ etc
  - Colour of each M&M in a box: Red, yellow etc
- Discrete Quantitative variables
  - for each orange tree in a grove, the number of oranges is counted: 50, 45,…., 70
  - for each day in May, the number of cars entering a college campus is counted: 122, 145, … 200
- Continuous Quantitative variables
  - time until a light bulb burns out: 1123.5 hrs
  - GPA of every student in the class: 11.45, 10.56, 7.80,…

EXERCISE

1.2)Identify each variable as quantitative or qualitative:

a) Amount of time it takes to assemble a simple puzzle
b) Number of students in the first-grade classroom
c) Rating of a newly elected politician (excellent, good, fair, poor)

EXERCISE

1.3) Identify the following quantitative variables as discrete or continuous:

a) Population of a particular area of Canada
b) Weight of newspapers recovered fro recycling on a single day
c) Time to complete a sociology exam
d) Number of consumers in a poll of 1000 who consider nutritional labelling on food products to be important

GRAPHS FOR CATEGORICAL DATA

- Use a data distribution to describe
  - What values of the variables have been measured
  - How often each value has occurred
    - Frequency = number of measurements in each category
    - Relative Frequency = Frequency/*n*
      (where *n* = sample size)
    - Percentage = Relative Frequency *100
- Sum of frequencies is n, Sum of relative frequency is 1, and sum of percentage is 100.

GRAPHS FOR CATEGORICAL DATA

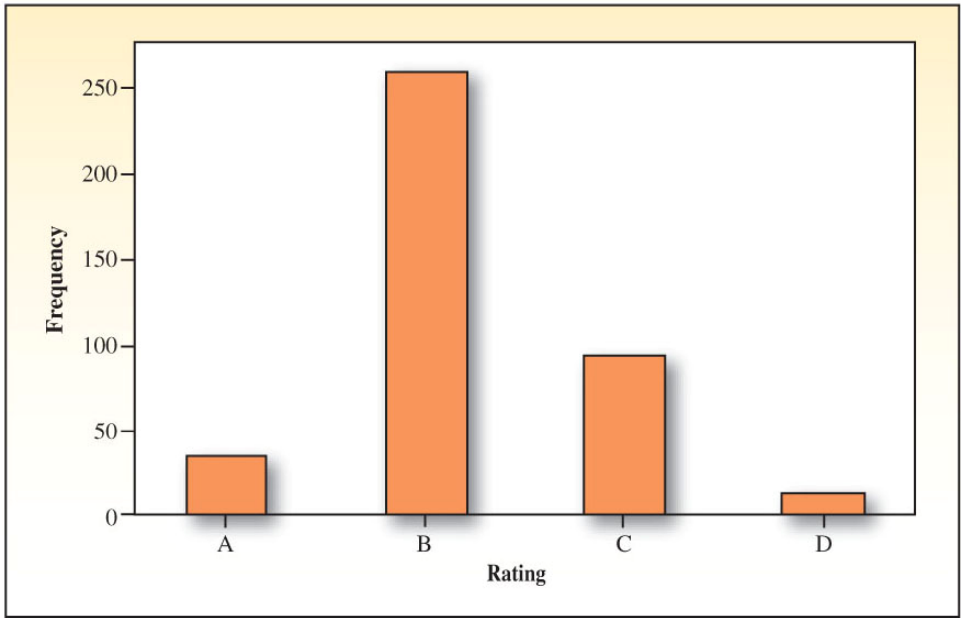
- The categories for qualitative variable should be chosen so that a measurement will belong to only one category, and each measurement has a category to which it can be assigned.
- Use **pie** chart or **bar** chart to display the data
- A bar chart in which the bars are ordered from largest to smallest is called a **Pareto** chart

EXAMPLE

- In a survey concerning public education, 400 school administrators were asked to rate the quality of education in Canada

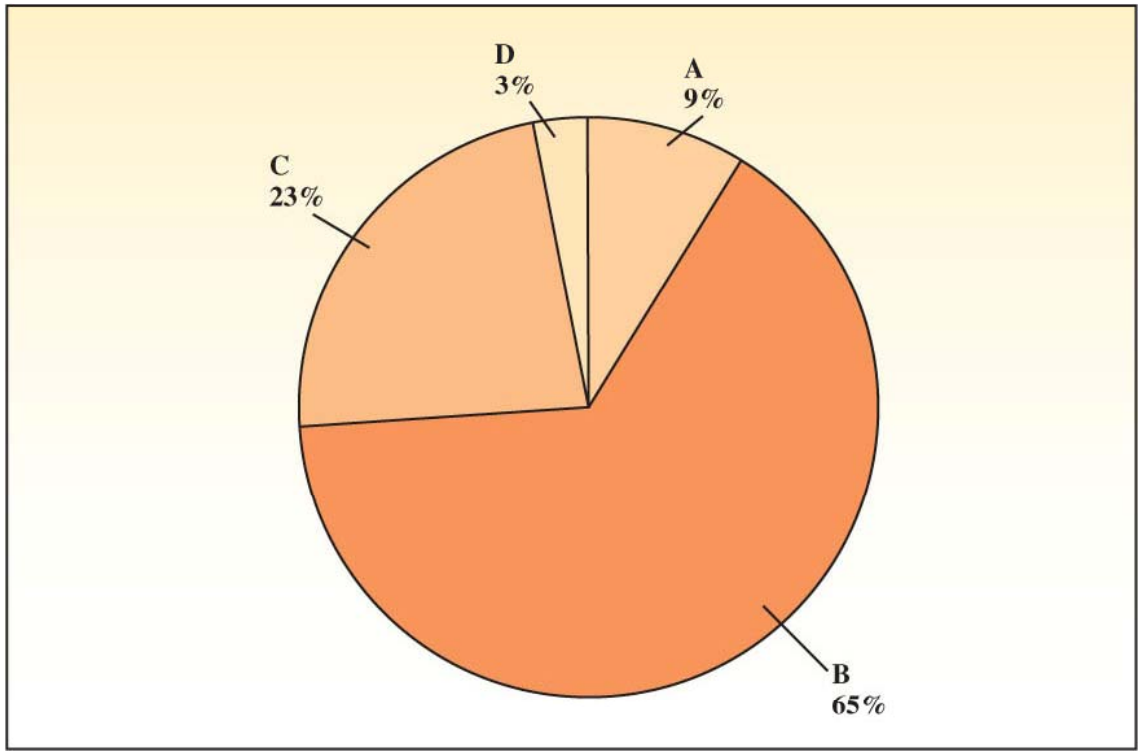
Rating	Frequency	Relative Frequency	Percent	Angle
A	35	35/400 = 0.09	9	$0.09 \times 360 = 32.4^\circ$
B	260	260/400 = 0.65	65	234.0°
C	93	93/400 = 0.23	23	82.8°
D	12	12/400 = 0.03	3	10.8°
Total	400	1.00	100%	360°

BAR CHART



Pie Chart

- Angle = Relative Frequency * 360



EXERCISE

1.4)A manufacturer of jeans has plants in Quebec (QC), Ontario (ON), and Manitoba (MB). A group of 25 pairs of jeans is randomly selected from the computerized database, and the province in which each is produced is recorded:

ON	QC	QC	MB	ON
ON	ON	MB	MB	MB
QC	QC	ON	QC	MB
ON	QC	MB	MB	MB
ON	QC	QC	ON	ON

EXERCISE-CONT'D

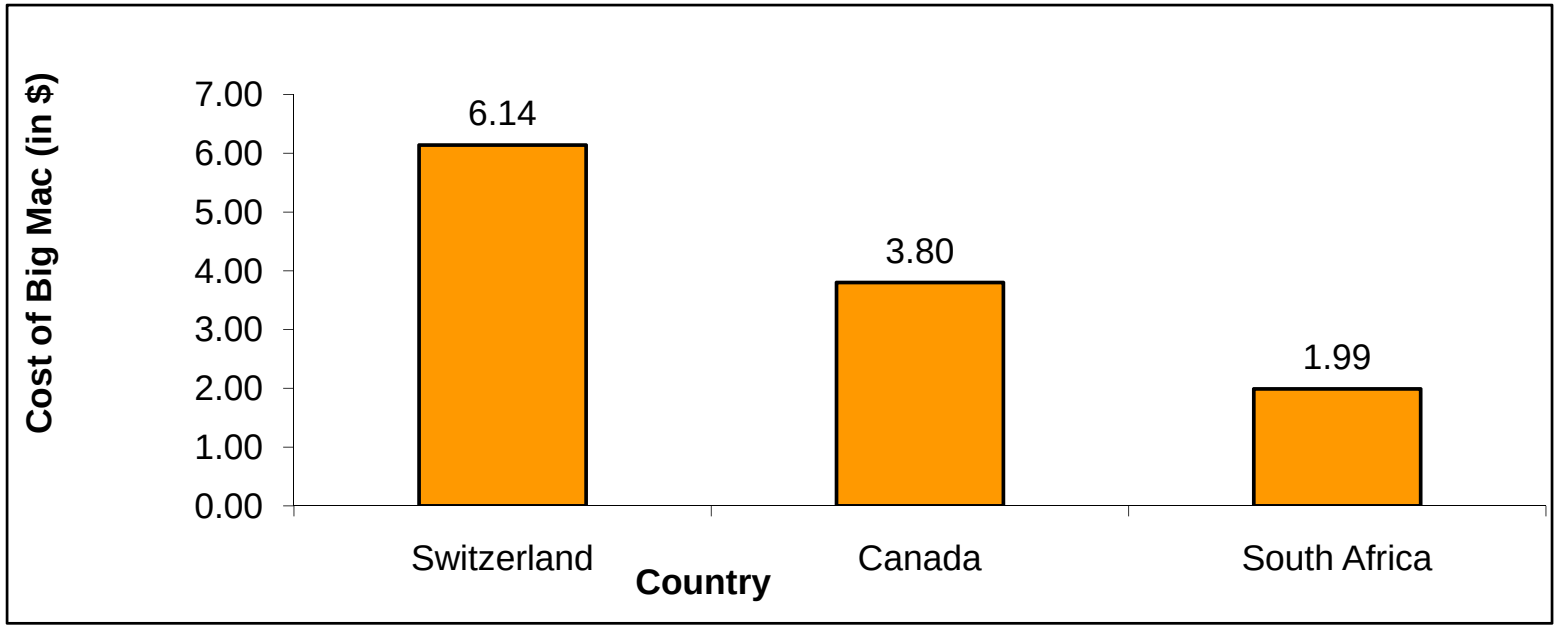
a) What is the experimental unit?
b) What is the variable being measured? Is it qualitative or quantitative?
c) Construct a pie chart to describe the data.
d) Construct a bar chart to describe the data.
e) What proportion of the jeans are made in Quebec?
f) What province produced the most jeans in the group?
g) If you want to find out whether the three plants produced equal numbers of jeans, or whether one produced more jeans than the others, how can you use the charts from parts c and d to help you? What conclusions can you draw from these data?

GRAPHS FOR QUANTITATIVE DATA

- When information is collected for a quantitative variable measured on different segments of the population or for different categories of classification, pie or bar chart can be used to display the data

Bar Chart for Quantitative Data

- Example: In July 2007, a Big Mac cost $6.14 in Switzerland, $3.80 in Canada, and $1.99 in South Africa. Price of big Mac in three different countries

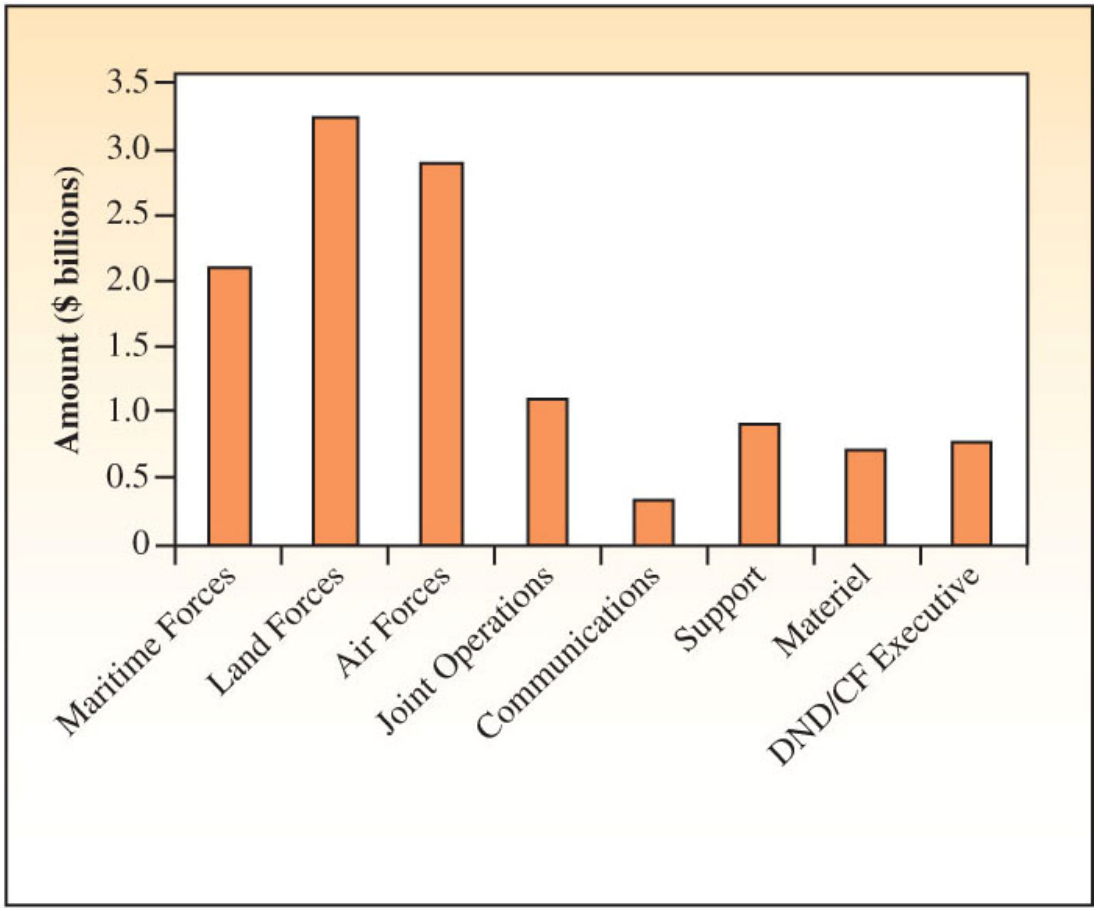


EXAMPLE

Category	Expenses ($billions)	Relative Frequency	Angle
Maritime Forces	2.05	0.18	64.8
Land Forces	3.18	0.27	97.2
Air Forces	2.83	0.24	86.4
Joint operation and civil emergency preparedness	1.09	0.08	28.8
Communications and information management	0.30	0.03	10.8
Support in the personal function	0.86	0.07	25.2
Material, infrastructure and environment support	0.75	0.06	21.6
DND/CF Executive	0.79	0.07	25.2
Total	11.85	1	360

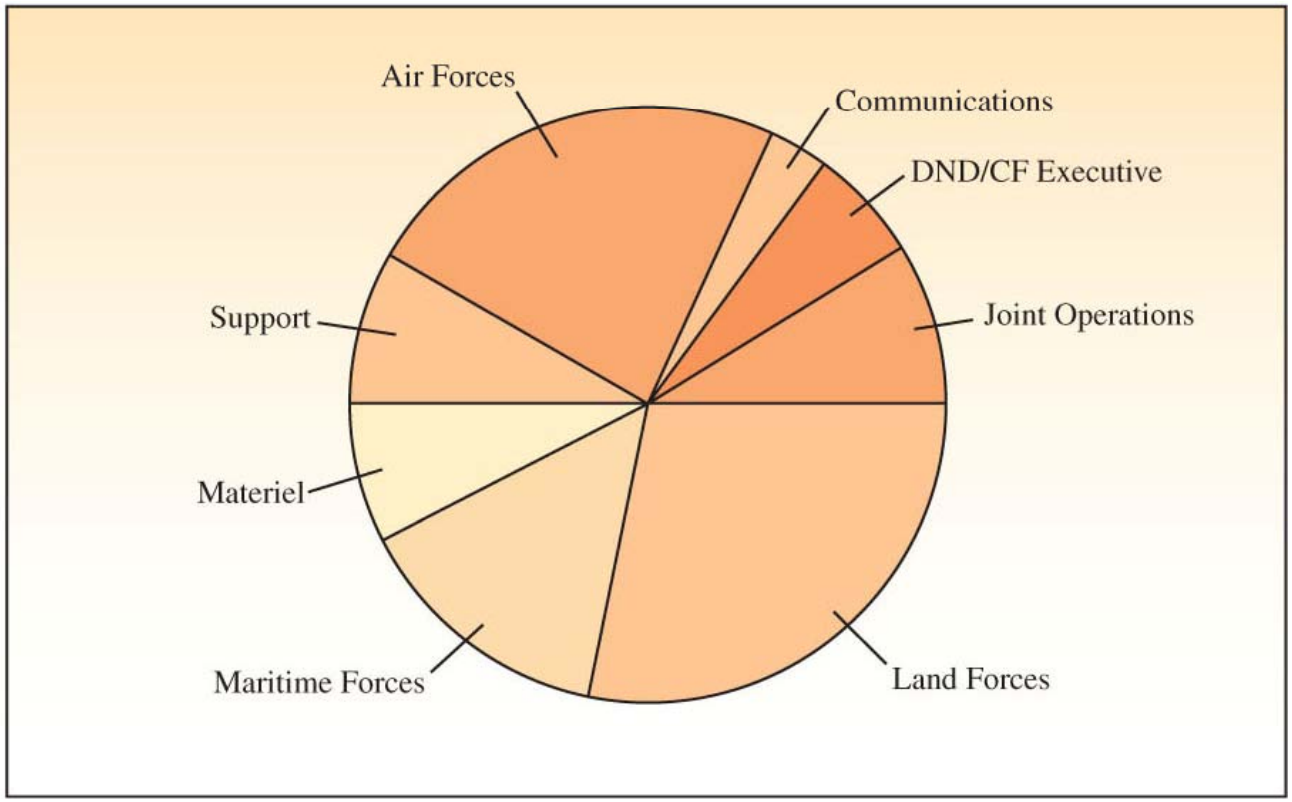
BAR CHART

Category of expenditure by Canadian Defence



Pie Chart

Category of expenditure by Canadian Defence

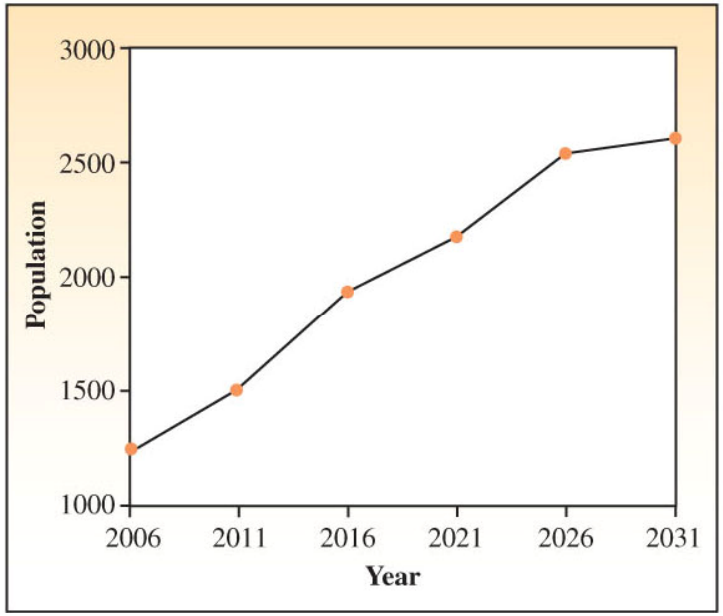


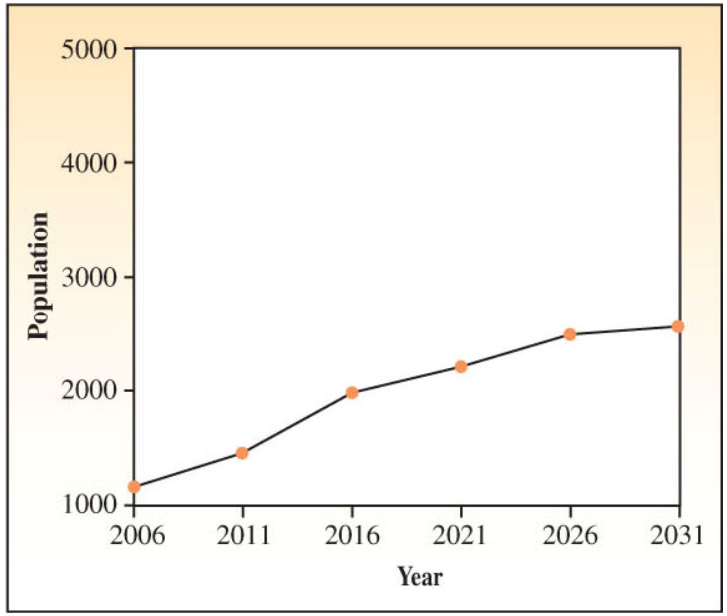
Line Chart

- When a quantitative variable is recorded over time at equally spaced intervals (such as daily, weekly, monthly, quarterly, or yearly), the data set forms a time series.
- Time series data most effectively presented on a **line** chart with time as the horizontal axis
- The trend observed in the line chart can be used to make accurate predictions for the immediate future

LINE CHART

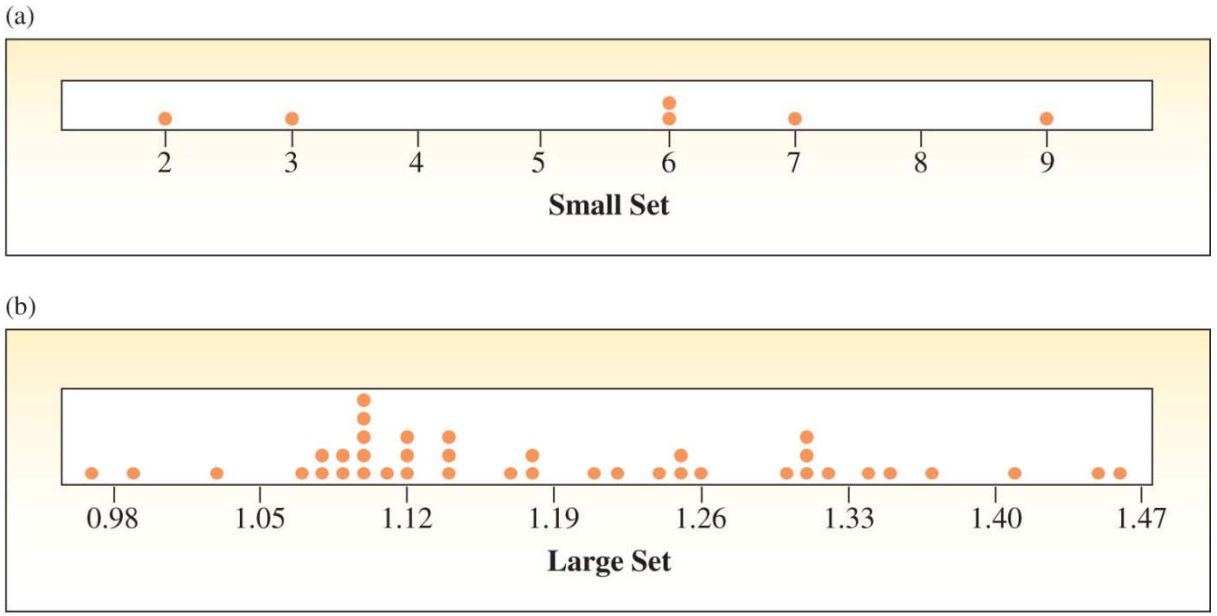
- Line chart for quantitative variable population measured over six time intervals shows steadily increasing population





Dot Plots

- The simplest graph for quantitative data, dot plots plot the measurements as points on a horizontal axis, stacking the points that duplicate existing points. (a) 2, 6, 9, 3, 7, 6



Stem and Leaf Plots

- A simple graph for quantitative data that uses the actual numerical values of each data point.
- How to construct Stem and Leaf Plot
  - Divide each measurement into two parts: the Stem and the leaf
  - List the stems in a column with a vertical line to their right
  - For each measurement, record the leaf portion in the same row as its corresponding stem
  - Order the leaves from lowest to highest in each stem
  - Provide a key to the stem and leaf

EXERCISE

1.5)The prices ($) of 19 brands of walking shoes: 90, 70, 70, 70, 75, 70, 65, 68, 60, 74, 70, 95, 75 , 70, 68, 65, 40, 65, 70. Plot the stem and leaf plot

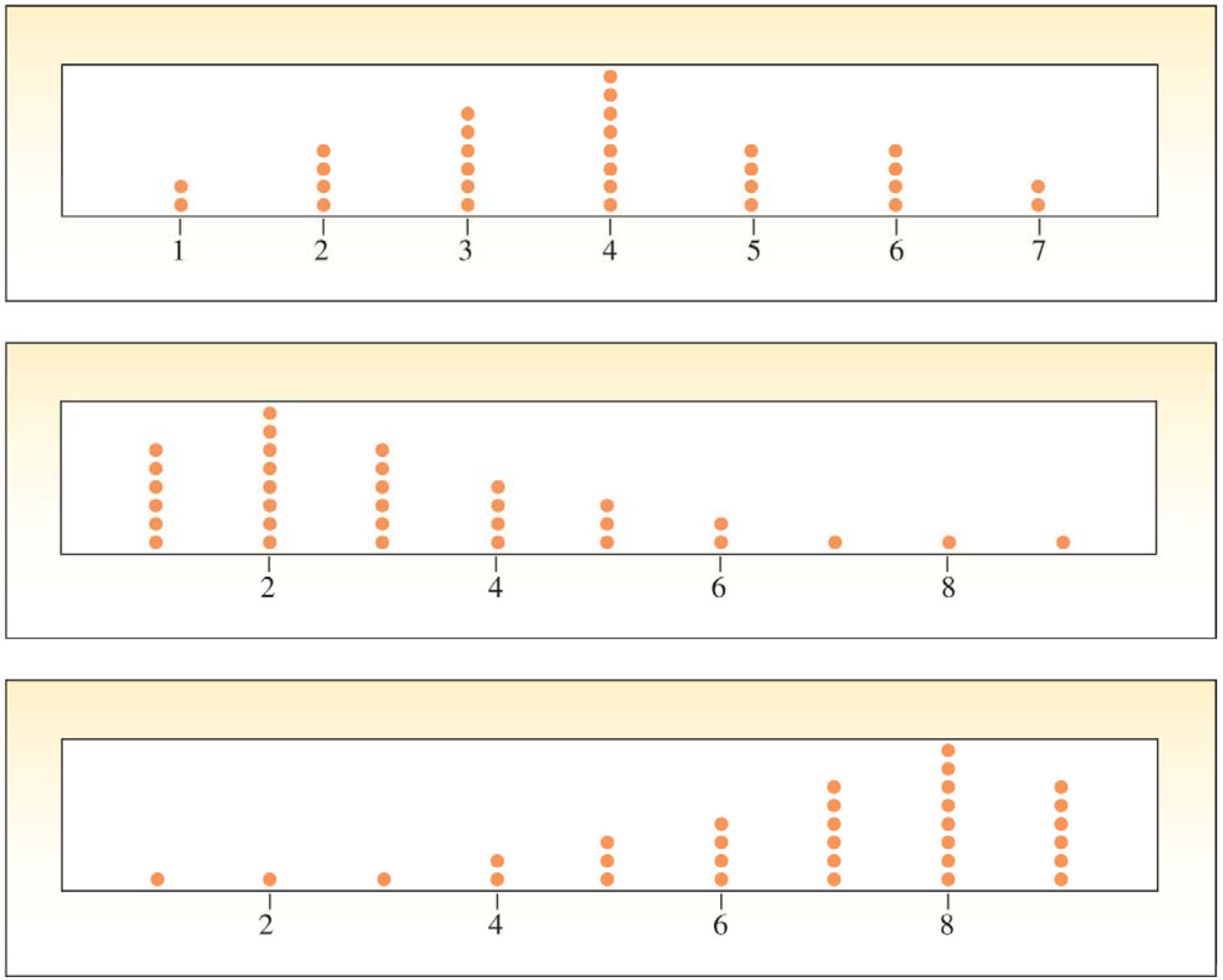
Interpreting Graphs: Location and Spread

- Information to look for graph of a set of data
  - Horizontal and vertical **scales**
  - **Location** of the distribution: where on the horizontal axis is the centre of the distribution
  - Examine the **shape** of the distribution: one or more peak (most frequently occurring category). Are there an approximately equal number of measurements to the left and right of the peak?
  - Look for any unusual measurements or **outliers**
- Compare graphs created for two data sets based their scales of measurement, locations, shapes, and look for outliers

SHAPES OF GRAPHS

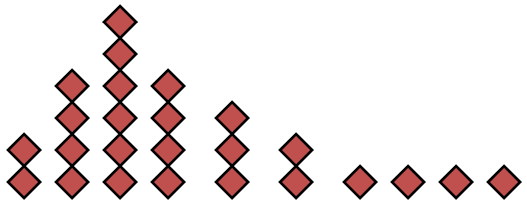
- **Symmetric/mound shaped:** If left and right side of distribution, when divided at the middle value, form mirror images
- **Skewed to the right:** if a greater proportion of the measurements lie to the right of the peak value. Contain a few unusually large measurements
- **Skewed to the left:** if a greater proportion of the measurements lie to the left of the peak value. Contain a few unusually small measurements
- **Unimodal** if it has one peak and Bimodal if it has two peaks.

Plots of Different Shapes

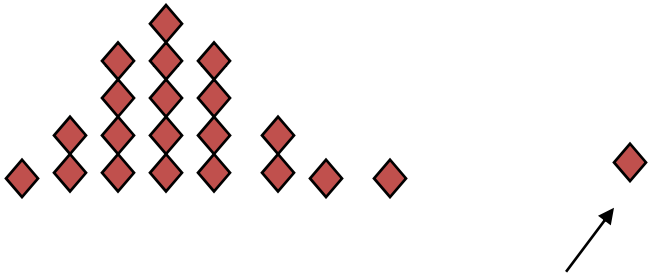


OUTLIERS AND BIMODAL

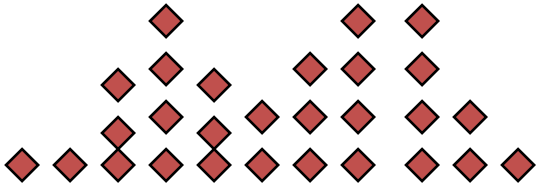
- Are there any unusual measurements that stand out in the data set



No Outliers



Outlier



Bimodal: two local peaks

RELATIVE FREQUENCY HISTOGRAMS

- A relative frequency histogram resembles a bar chart, but it is used to graph quantitative variables.
- A bar graph in which the height of the bar shows "how often" measurements fall in a particular class or subinterval.
- Choose number classes (usually 5 -12)
- Calculate the class width by dividing the difference between the largest and smallest values by the number of classes. Round the width up to a convenient number

Relative Frequency Histograms

- Discrete data – one class for each integer value taken on by data
- Large number of integer values or continuous data – group them into classes
- Locate the class boundaries. Lowest class must include the smallest measurement. Then add the remaining classes using the left inclusion method
- Calculate the frequencies and relative frequencies for each class. Construct the histogram

STATISTICAL TABLE TO CONSTRUCT HISTOGRAM

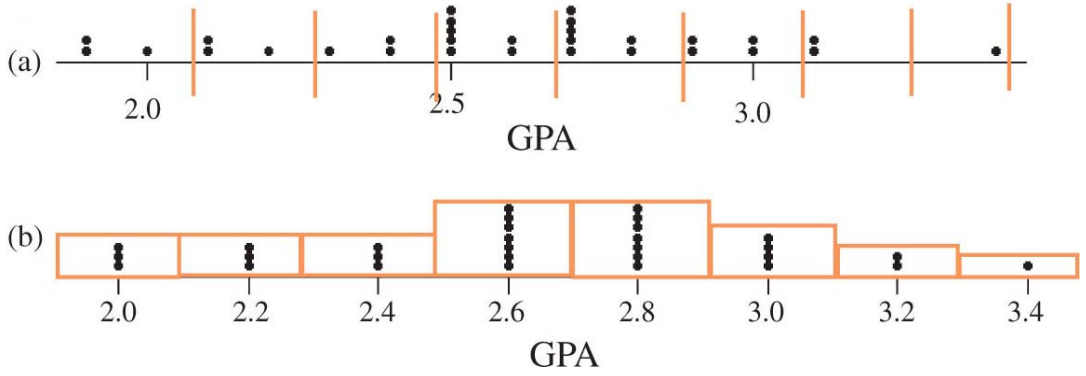
- Data (30 observations) : 2.0, 3.1, **1.9**, 2.5, 1.9, 2.3, 2.6, 3.1, 2.5, 2.1,2.9,3.0, 2.7, 2.5, 2.4, 2.7, 2.5, 2.4, 3.0, **3.4**, 2.6,2.8, 2.5, 2.7, 2.9, 2.7, 2.8, 2.2, 2.7, 2.1.
- Class width = (3.4-1.9)/8 = 0.1875≈ 0.2

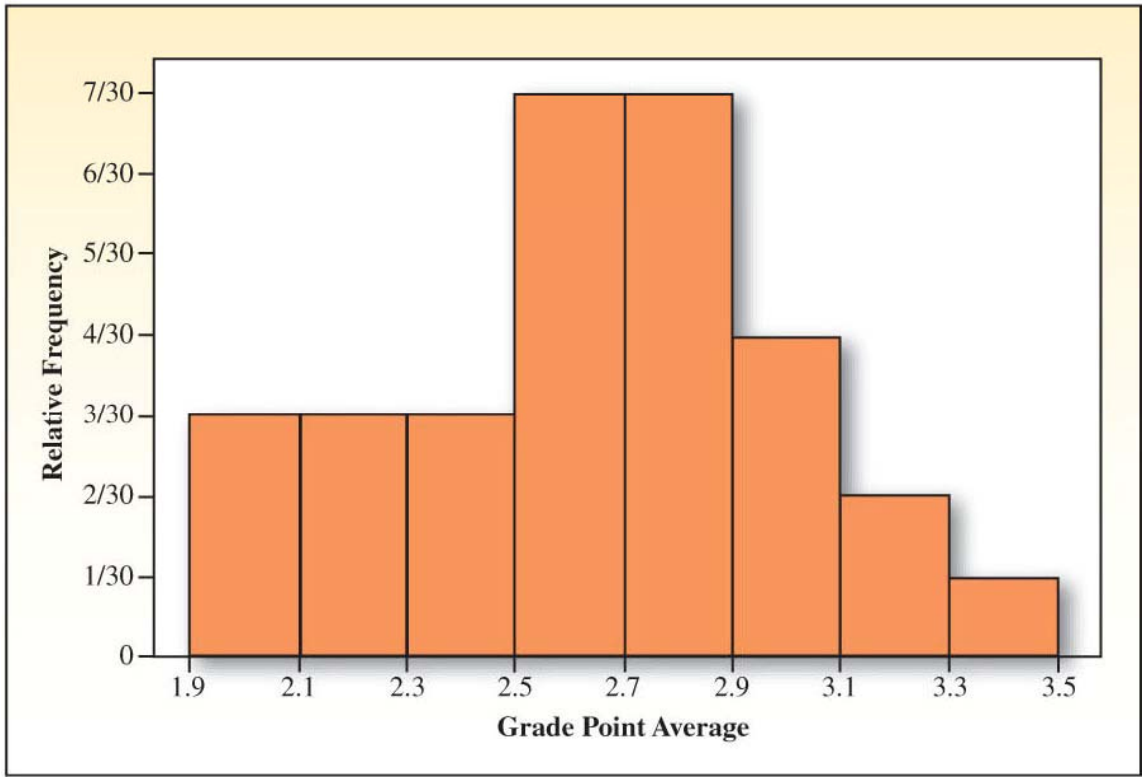
STATISTICAL TABLE TO CONSTRUCT HISTOGRAM

- Table

Class	Class Boundary	Frequency	Relative Frequency
1	1.9 to<2.1	3	3/30
2	2.1 to <2.3	3	3/30
3	2.3 to <2.5	3	3/30
4	2.5 to < 2.7	7	7/30
5	2.7 to < 2.9	7	7/30
6	2.9 to < 3.1	4	4/30
7	3.1 to < 3.3	2	2/30
8	3.3 to < 3.5	1	1/30

RELATIVE FREQUENCY HISTOGRAMS





EXERCISE

1.6) Consider this set of data:

4.5	3.2	3.5	3.9	3.5	3.9
4.3	4.8	3.6	3.3	4.3	4.2
3.9	3.7	4.3	4.4	3.4	4.2
4.4	4.0	3.6	3.5	3.9	4.0

a) Construct a stem and leaf plot by using the leading digit as the stem and frequency histogram with two classes

b) Construct a stem and leaf plot by using each leading digit twice. Does this technique improve the presentation of data? Explain.

c) Construct a stem and leaf plot by using each leading digit five times.

SUMMARY

- Data
  - Experimental Units, Variables, measurements
  - Samples and Populations
  - Univariate, bivariate, and multivariate data
- Types of Variables
  - Qualitative of categorical
  - Quantitative
    - Discrete
    - Continuous

SUMMARY

- Graphs for Univariate Data Distributions
  - Qualitative or categorical data
    - Pie charts
    - Bar charts
  - Quantitative data
    - Pie, bar charts, Pareto chart, Line chart
    - Dot plots
    - Stem and leaf plots
    - Relative frequency histograms
  - Describing data distributions
    - Shapes – Symmetric, skewed left, skewed right, unimodal, bimodal
    - Proportion of measurements in certain intervals
    - Outliers