Catalogue no. 12-004-X

Statistics: Power from Data!

Release date: September 2, 2021

Statistics Canada Statistique Canada

Canada

How to obtain more information

For information about this product or the wide range of services and data available from Statistics Canada, visit our website, www.statcan.gc.ca.

You can also contact us by

Email at infostats@statcan.gc.ca

Telephone, from Monday to Friday, 8:30 a.m. to 4:30 p.m., at the following numbers:

• Statistical Information Service	1-800-263-1136
• National telecommunications device for the hearing impaired	1-800-363-7629
• Fax line	1-514-283-9350

Depository Services Program

• Inquiries line	1-800-635-7943
• Fax line	1-800-565-7757

Standards of service to the public

Statistics Canada is committed to serving its clients in a prompt, reliable and courteous manner. To this end, Statistics Canada has developed standards of service that its employees observe. To obtain a copy of these service standards, please contact Statistics Canada toll-free at 1-800-263-1136. The service standards are also published on www.statcan.gc.ca under "Contact us" > "Standards of service to the public."

Note of appreciation

Canada owes the success of its statistical system to a long-standing partnership between Statistics Canada, the citizens of Canada, its businesses, governments and other institutions. Accurate and timely statistical information could not be produced without their continued co-operation and goodwill.

An HTML version is also available.

Cette publication est aussi disponible en français.

Table of contents

Statistics: Power from Data!

Statistics: Power from Data! is a training tool for students, teachers and the general population that will help them in getting the most from statistics. This resource aims to help readers:

- gain confidence in using statistical information,
- appreciate the importance of statistical information in today's society,
- make critical use of information that is presented to them.

These goals are at the heart of Statistics Canada's mission to assist Canadians with informed decision-making based on data.

The first section of this resource defines the concepts of data, statistics and statistical information, as well as data quality. The second section describes the types of data that can be used to produce statistical information. The third section details the processes involved in the production of statistical information, such as sampling, data collection, edit and imputation, estimation and record linkage. Sections 4 and 5 explain how to use descriptive statistics and data visualization to explore data. Each chapter is intended to be complete in itself, and contains exercises to help consolidate the understanding of the material.

It is important to note that the length and amount of detail of a section don't reflect the importance of its topic to the overall survey process, but rather the expected technical level of the target audience. This explains why the section on creating graphs is very detailed, while the one on estimation and weights is quite brief. Graphing is an appropriate tool for secondary schools, but the details of data estimation necessitate advanced knowledge in mathematical statistics that is usually acquired during postsecondary studies.

Finally, special thanks must be given to the Australian Bureau of Statistics for allowing Statistics Canada to use the second edition of their book **Statistics — A Powerful Edge!** as a basis in developing this online Canadian counterpart at the start of the 2000s and whose influence is still as important for this update to **Statistics: Power from Data!**

1 Data, statistical information and statistics

Our data-rich society is facing new and challenging problems sifting and interpreting an ever-growing body of information. More than ever before, governments, industry, and citizens need reliable statistical information to make better decisions, but the task of providing high-quality information when it is needed continues to grow in complexity.

The need for an informed society is one reason why the Canadian education system is developing curricula that emphasize the gathering, processing and presentation of data. However, before an information user can undertake such activities, it is important for them to have a sound understanding of the terms **data**, **statistical information** and **statistics**, as well as a means to evaluate their quality.

1.1 Definitions

Data, statistical information and statistics are closely related, but understanding the key differences between these concepts is important for anyone who needs to navigate the ever-rising ocean of information produced by modern society. Data are the raw materials for producing statistical information, of which statistics are a specific type.

Data

Data are facts, figures, observations, or recordings that can take the form of image, sound, text or physical measurements (ex: distance, weight, wave lengths). Data can be gathered and processed in order to form conclusions. Data can come from many sources and it can be split in two groups based on the form it takes: structured data and unstructured data.

Structured data are data that are organized into pre-defined items that each relates to a specific concept or data item. A set of data gathered using a questionnaire or other fillable form is a good example of structured data: the questions on a questionnaire represent separate, well-defined concepts. In the case of a closed question, the answer will fit in one of multiple pre-defined categories. For an open question, it may take the form of a text or numerical values. If an answer was recorded for each question, the data are complete. If not, there are missing values.

For example, consider how each column in table 1.1.1 on Canadian universities relates to a single, separate concept:

Table 1.1.1
Example of structured data

Name of institution	City	Province	Established	Number of students
Université Laval	Quebec	QC	1852	43,000
University of Waterloo	Waterloo	ON	1955	30,000
Dalhousie University	Halifax	NS	1818	18,000
Simon Fraser University	Burnaby	BC	1965	30,000

Each row includes the values for one observation unit for which information was collected. Rows are referred to as observations or records. Concepts presented in each column are often called variables. Data sets are groupings of data that have common definitions of observation units and variables.

In order to be processed and analyzed, structured data need to be compiled in a digital data structure that naturally aligns with pre-defined concepts or variables such as a spreadsheet, a database or a delimited text file. Data can then be read by a statistical software that allows the data user to transform and summarize the data, to perform mathematical operations on the data or to visualize them.

Unstructured data are any data that are not arranged according to a pre-defined model. To produce statistical information based on unstructured data, additional processing is needed to organize the information contained in the data. Table 1.1.2 presents examples of how text, images and sounds can be transformed into structured data that can be used for text analysis and for pattern and speech recognition.

Table 1.1.2
Transforming unstructured data into structured data

Unstructured data	Processing	Structured data
A text	Parsing, to split the text in a list of words; aggregation, to count how many times the same word occurs; use of dictionaries and rules to classify words.	A spreadsheet: on each row there is one distinct word, the three columns present the word, the number of occurrences and the category of the word.
An image	Assignment of RGB values to pixels; segmentation of the image into blocks of pixels based on red (R), green (G) and blue (B) components.	A database: each record is a group of pixels and the variables summarize the colour components in each group.
A record of someone's voice	Segmentation of record in distinct sounds; measure of duration and frequencies.	A list of segments with duration and frequencies.

With the increased use of computers and smartphones in all areas of our lives, a huge part of the digital data that is being created now is unstructured. Assessing the potential of this data and creating innovative ways of gathering, processing and analyzing it in order to produce valuable statistical information is one of the great challenges of the data revolution.

But what is the difference between statistical information and data?

Statistical information

Statistical information is data that has been recorded, classified, organized, related, or interpreted within a framework so that meaning emerges. Statistical information that is communicated to information users should help them understand the story told by the data and communicate to them the quality of the information that is presented. Statistical information can be presented in various formats: texts, tables, graphs, infographics, videos, or even databases.

Many examples of statistical information produced at Statistics Canada will be presented in the next page, but it is first important to understand one major part of the process of producing statistical information from data: the use of statistics!

Statistics

In general, statistics relate to numerical data; in fact, the term “statistics” can refer to the science of dealing with numerical data itself. Statistics are also a type of information obtained through mathematical operations on data. Above all, statistics aim to provide useful information by means of numbers.

The most commonly used statistics to report statistical information are called descriptive statistics. For numeric variables, measures of central tendency provide the value that is the most representative of the units found in a data set. Measures of dispersion describe the spread of the data around the central tendency. For categorical variables, frequency distributions are used to summarize the data. Proportions, ratios and rates are also useful statistics to analyze the data.

When each row in a data set displays statistics that summarize the information for many units of observation, these data are called aggregate data. Inversely, when each row displays the information for a single unit of observation, the data are referred to as microdata.

1.2 Examples of statistical information

As you will see, statistical information can be presented in a variety of ways such as texts, tables, visualizations, or infographics.

Quarterly retail sales

An example of statistical information that can be used for decision-making is given below. At Statistics Canada, release of new statistical information on key economic indicators is often done through a short-written communication in The Daily.

> In the second quarter of 2019, retail sales in Canada reached $163.3 billion, up 1.4% from the same quarter of 2018. Sales were up in 13 of 19 commodity groupings for the second quarter of 2019.
>
> The largest increase in dollar terms came from food, which posted a year-over-year growth of 3.5%. The majority of this gain was attributable to higher sales of fresh food (+3.4%), led by growth in sales of fresh fruit and vegetables (+5.4%). Sales of packaged food dry goods increased 4.4%, while sales of frozen food grew 1.1%. Sales of soft drinks and alcoholic beverages rose 2.3% from the previous year, largely because of higher sales of alcoholic beverages (+2.6%).
>
> Sales of motor vehicle parts, accessories and supplies grew 5.0% in the second quarter, largely on higher sales of motor vehicle parts and accessories (+6.0%) and new motor vehicle tires (+3.6%).
>
> Meanwhile, sales of motor vehicles also rose 0.6% in the same period as a result of higher sales of used motor vehicles (+6.7%). Receipts for new motor vehicles declined 2.5%, led by lower sales of new passenger automobiles (-14.9%).
>
> Hardware, tools, renovation and lawn and garden product sales were up 1.9% in the second quarter. The largest contributor to the increase was home lawn and garden equipment and supplies (+8.2%).
>
> Sales of automotive and household fuels were down 2.7% compared with the second quarter of 2018. This decline was largely attributable to lower sales of automotive fuels (-2.6%).
>
> Sales of cannabis products in the second quarter of 2019 totalled $251.8 million. Sales of dried cannabis flowering tops were $222.4 million, while sales of cannabis oils were $29.0 million.
>
> **Source:** Statistics Canada, The Daily, 15/10/2019

1911 Census occupations

This is a table of statistical information about the types of occupations available in Canada in the last century. It shows the number of Canadians who had particular occupations at the time of the 1911 Census. Note how some occupations were referred to at that time!

Table 1.2.1
Selected occupations from the 1911 Canada Census

Occupation	Males	Females
Bridge and gate tenders	436	2
Char-workers	12	4,700
Launderers and laundresses	588	282
Hotel keepers	3,102	848
Undertakers	43	0
Gardeners	469	18
Coachmen and grooms	418	0
Sailors and seamen	16,347	0
Match makers	72	178
Nurses	124	5,476
Stenographers and typists	1,603	9,754
Actors and theatrical employees	2,410	432
Musicians and teachers of music	2,001	3,574
Hucksters and peddlers	3,135	113
Total	**30,760**	**25,377**

0 true zero or a value rounded to zero

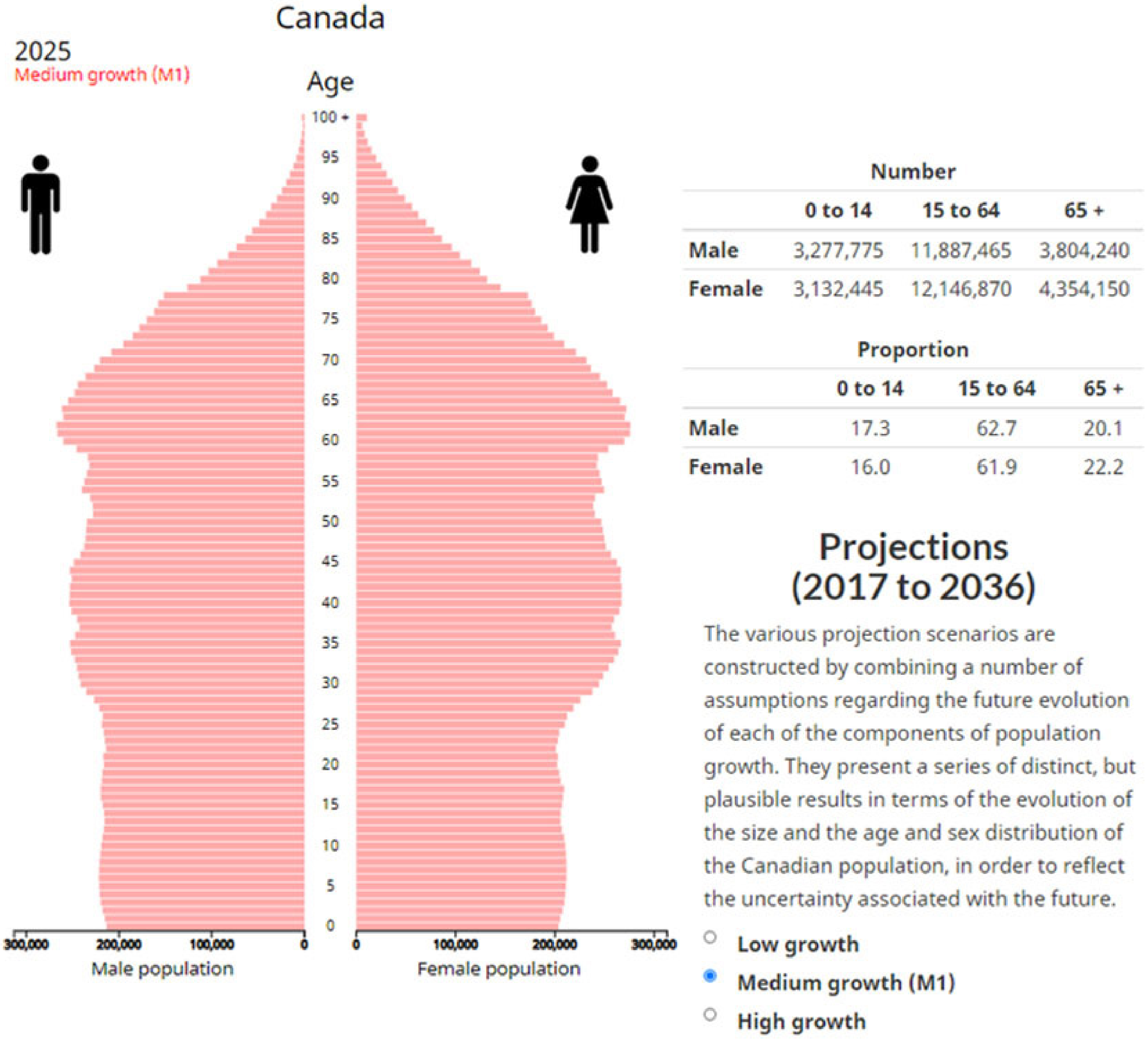
Population pyramids

Figure 1.2.1 is an example of a dynamic age—sex pyramid of Canadian population projections for 2025, based on an assumption of medium growth. It consists of two horizontal histograms, placed side by side, indicating the number of persons in each single year of age. Men are shown on the left histogram and women are shown on the right, as stated by convention.

View the product Historical Age Pyramid for other years and other projection scenarios.

Age–sex pyramids are commonly used to present statistical information on the age and sex composition of a population. This chart clearly shows the aging "Baby Boomers."

Figure 1.2.1
Population pyramid of the Canadian population projected in 2025



Number

	0 to 14	15 to 64	65 +
Male	3,277,775	11,887,465	3,804,240
Female	3,132,445	12,146,870	4,354,150

Proportion

	0 to 14	15 to 64	65 +
Male	17.3	62.7	20.1
Female	16.0	61.9	22.2

Projections (2017 to 2036)

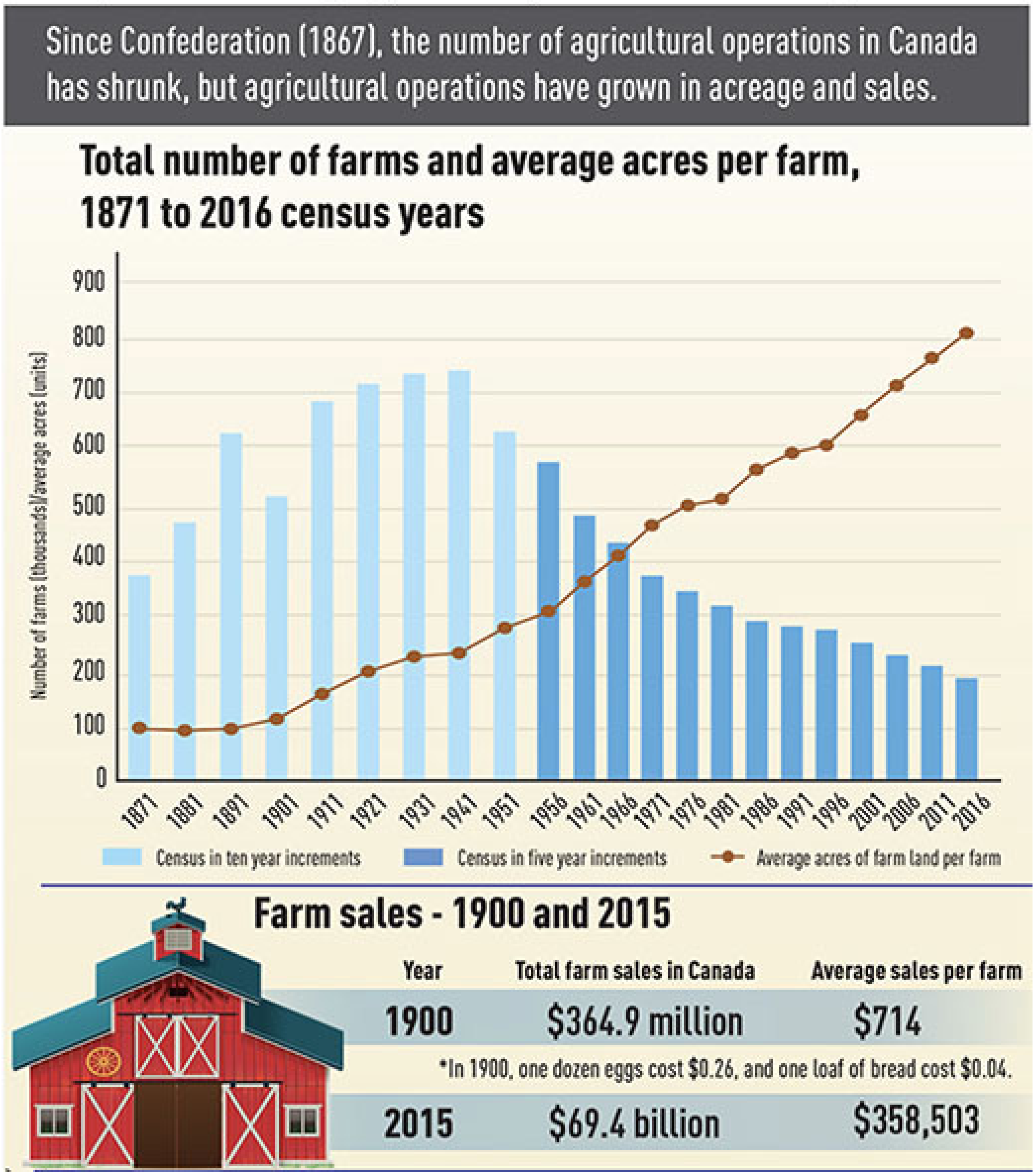
The various projection scenarios are constructed by combining a number of assumptions regarding the future evolution of each of the components of population growth. They present a series of distinct, but plausible results in terms of the evolution of the size and the age and sex distribution of the Canadian population, in order to reflect the uncertainty associated with the future.

○ Low growth
● Medium growth (M1)
○ High growth

Infographics

Infographics combine statistics, visualizations (such as graphs or pictographs) and text to communicate a large amount of statistical information in a compact, visually appealing way. Figure 1.2.2 below is a clip from an infographic entitled "150 years of Canadian agriculture." It consists of a chart showing the time series for the total number of farms and the average acres by farm as well as a table of the total farm sales in 1900 and 2015.

Figure 1.2.2
Excerpt from the infographic "150 years of Canadian agriculture"



Since Confederation (1867), the number of agricultural operations in Canada has shrunk, but agricultural operations have grown in acreage and sales.

Total number of farms and average acres per farm, 1871 to 2016 census years

Farm sales - 1900 and 2015

Year	Total farm sales in Canada	Average sales per farm
1900	$364.9 million	$714
*In 1900, one dozen eggs cost $0.26, and one loaf of bread cost $0.04.		
2015	$69.4 billion	$358,503

1.3 Data quality

Generally speaking, statistical information is evaluated in terms of its “fitness for use” - that is, the extent to which the statistical information can be relied upon to fulfill the user’s information needs. At Statistics Canada, fitness for use is considered along six quality dimensions.

Quality dimensions and examples of questions to ask

Relevance

Does the statistical information matter to Canadians?

- Does it fill a data gap?
- Is it useful in building policies?
- Does it aid in long-term planning?
- Can it promote new initiatives?

Accessibility

Can users access the statistical information?

- Is it easy to access?
- Is it affordable?
- Is it organized and easy to locate?
- Can users who encounter difficulties in accessing information request assistance?

Accuracy

Is the statistical information representative of the targeted measurement?

- Does it cover the required population and period of reference?
- Are there known sources of under-coverage?
- Are methods transparent?
- Was the information produced without external influence?

Timeliness

Is the lag between the period of reference and the availability of the statistical information acceptable?

- Is the statistical information available when it is the most needed?
- Are you willing to accept lower accuracy to get the data faster?

Interpretability

The metadata is the information that provides context to data.

- Is metadata available and complete?
- Are they useful?
- Are they reliable?
- Are they available at the same time as the statistical information to which they pertain?

Coherence

Is the statistical information consistent over time, between region and across sub-populations?

- Does it use standard concepts and classifications?
- Was it produced using methods that are common to other statistical products?
- Is it comparable to previously released statistical information?

Often, a compromise is necessary between quality dimensions. For example, the need for timeliness can impact accuracy: publishing statistical information quickly reduces the time available for ensuring the accuracy of the information.

Besides data quality, it is also important to consider the ethics of data that are collected and of processes used to produce the statistical information. Ethically sourced data are collected in a transparent manner and used in a meaningful way that will not cause harm to the respondents.

1.4 Exercises

1. Identify which items in the list are structured data and which items are unstructured data.

   a. A bank statement
   b. An email
   c. A grocery store flyer
   d. A school report
   e. Results from a web search engine

2. Find which concept is associated with each definition.

   a. Some observation units don't have a value for one variable in a data set.
   b. Data that have been classified and interpreted.
   c. The most representative value of the units in data sets.
   d. A software that allows users to make mathematical operation on numeric data.
   e. The lag between the period of reference and the availability of the information.
   f. A data set in which a record represents only one observation unit.
   g. A visualization that combines images with statistical information to tell the story of the data.

3. Identify which of the six quality dimensions is problematic in each of the situations listed below.

   a. You found the perfect data set for a school work, but there is a cost to be able to get the data.
   b. You want to calculate the average age of people in your school, but you only have the age of the students.
   c. You want to explore a data set, but you don't know what each variable means because the names of the variables are not explicit.
   d. You made a survey in your class to know the level of physical activity of your classmates, but some replied with a number of steps and some with a distance in kilometres. For this reason, it is hard to identify those who do the most physical activity in a week.

4. True or false?

   a. Statistics are raw materials from which the statistical information is created.
   b. In 2025, there will be fewer 15-year-old people than 45-year-old people in Canada.
   c. In the second quarter of 2019, retail sales had increased in all commodity groupings compared to the previous year.
   d. The goal of statistics is to provide useful information using numbers.

1.5 Answers

1. a. Structured
   b. Unstructured
   c. Unstructured
   d. Structured
   e. Unstructured

2. a. Missing value
   b. Statistical information
   c. Measure of central tendency
   d. Statistical software
   e. Timeliness
   f. Microdata
   g. Infographic

3. a. Accessibility
   b. Accuracy
   c. Interpretability
   d. Coherence

4. a. False
   b. True
   c. False
   d. True

2 Sources of data

Section 1 noted that data are like raw materials used as inputs to the production of statistics and statistical information. But where exactly do these raw materials come from, and who is interested in processing it to produce statistical information or to use that statistical information to make decisions? In this section we first discuss data providers and data users, which leads to further exploration of different types of data that can be used to produce statistical information, as well as the advantages and disadvantages of each type.

2.1 Data providers and data users

Data providers are individuals and organizations that collect and process data for a variety of purposes. They make these data accessible to data users who, in turn, use it to produce statistical information. Data users can be part of the same organization as the data providers or be a third party. Depending on the objective of the data collection, the source of data is said to be primary or secondary.

Data from a primary source is collected for the purpose of producing statistics and statistical information. Researchers, enterprises and government agencies are the main primary data sources. Researchers and special interest groups may obtain funding to study social, economic and scientific issues which necessitate collecting new data if there is an information gap, that is, if the data needed to study these issues is not readily available elsewhere. Enterprises may collect data for statistical purposes as part of their main mission (survey firms, market research) or for internal uses (survey of employees, for instance). Government agencies, like National Statistical Offices such as Statistics Canada, have mandates to collect information about different aspects of populations and society to help governments make decisions and build public policies.

Data from a secondary source is collected for a purpose other than producing statistical information. Any individual or organization can collect a large quantity of data for different uses, such as paying employees, maintaining inventories, and improving smart phone applications. Not all data from secondary sources is fit for producing statistical information, however, data users are increasingly interested in using some of it to produce statistical information. When data users choose to use secondary data, it is important to assess the data providers and their motives when they collected the data. This will help ensure that the data fit to use.

Groups and organizations that use data, statistical information and statistics include the following:

- **Governments:** Federal, provincial and local governments need information on the population and the economy, among other things, to help them develop, implement and monitor socio-economic and environmental programs and other functions of government, such as licensing and regulation. It helps governments make decisions on issues such as where to build hospitals, locate services, or how much money to raise through taxation. It also allows the public, opposition politicians and interest groups to measure a government's performance in decision making and to hold it accountable if it does not meet the criteria.
- **Businesses:** Canadian businesses require information about the local, provincial and national economy, the unmet needs of a population, and trends in society. Data helps them make decisions about employing people, marketing their products and opening new offices, warehouses and factories. Data are also required for businesses to carry out their operations, such as billing, inventory and supply management.
- **Community groups:** These organizations need information about a wide variety of subjects, such as health and population distribution of Indigenous people, or the number and location of Canadian immigrants who require English or French language skills. Sporting clubs may want information about attendance at games or the number of young people in their local area.
- **Academics and researchers:** Data is required by those carrying out studies and other analyses, in a variety of roles. Data may be used in planning research work (e.g. in what community to conduct a study) or to back up research claims and other hypotheses (e.g. do historic data support a correlation between higher average temperatures and increased flooding).
- **Individuals:** Everyone, from students to pensioners, needs some form of information at some time during their lives. The information may be used to complete an essay, a major project or simply to satisfy one's curiosity.

The examples above illustrate that the needs of data and information users span from pre-processed data to finished statistical information products, and everything in between. That is, information users may also be information producers, to varying degrees.

2.2 Type of data

There are many methods to collect data, but agencies like Statistics Canada primarily use three methods of data collection: censuses, sample surveys, and administrative data. Each has advantages and disadvantages that will be presented in this section. Then other methods of data gathering will be described.

Census

In general, census refers to data collection about every unit in a group or population. If you collected data about the height of everyone in your class, that would be regarded as a census of your class. Censuses are often used not only to collect data about all units of a population, but also to list and count all units of a population. If you wanted to know how many people live in your street, you would need to list all of the dwellings in the street and then all people living at each of these dwellings. As you do so, you could decide to collect other information, such as age, sex and mother tongue. That would allow you to count how many men, women and children are living in your street. So, a census would be a straightforward means to count the number of units and to produce statistics on their characteristics as well.

Here are some advantages and disadvantages of using a census:

Advantages (+)

No sampling variability: There is no sampling variability attributed to the statistics produced from a census because they are calculated using the entire population.

High level of details: With a census, you would be able to produce statistics for small sub-groups of the population, as long as you collected the right classification variables.

Direct estimation of counts: A census allows for the direct estimation of the population counts, although some adjustments may be considered for units that couldn’t be reached.

Disadvantages (-)

High cost: Conducting a census of a large population can be very expensive.

Timeliness: A census generally takes longer to conduct than a sample survey, which means the lag between the reference date and the release of results could be much larger.

High response burden: Information needs to be received from every member of the population.

Less control on quality: If the size of the population is much larger than the sample size of a survey and resources are limited, it is possible that some compromises on quality control will be necessary. For example, it is possible that only part of the non-respondents will be reached for nonresponse follow-up.

Less detailed information: Due to cost, response burden and scale of activities needed to carry a census in a large population, the variables that can be measured for each unit are sometimes limited to a short list of identification and classification variables.

Sample survey

A survey is any activity to collect information in an organized and methodical manner about the characteristics of the units of a population. At Statistics Canada, surveys use well-defined concepts, methods and procedures that will be described in the third section of this resource. A census can be seen as a type of survey, but the term **survey** is more often used to refer to a **sample survey**, which means a survey where the information is collected for some units of the target population only. If you collected data about the height of 10 students in a class of

30, that would be a sample survey of the class rather than a census. But ideally, you would want to select them randomly to make sure the 10 students are representative of all students in the class.

Here are some advantages and disadvantages of using a sample survey compared to using a census:

Advantages (+)

Lower cost: A sample survey costs less than a census because data is collected from only part of a population.

Faster results: Results are obtained far more quickly as fewer units need to be contacted and fewer data need to be processed.

Lower response burden: Fewer people have to respond to the survey.

More control on quality: The smaller scale of operations allows for better monitoring and quality control.

Disadvantages (-)

Sampling variability: If you selected multiple samples from the same population and computed statistics for each of those samples, the results would be a bit different from one sample to another. This source of uncertainty must be taken account in the estimation of statistics from a sample survey.

Lower level of details: The sample may not be large enough to produce information about sub-groups of the population or small geographical areas.

Administrative data

Administrative data is collected as a result of an organization’s day-to-day operations. Examples include data on births, deaths, tax, car registrations and transactional data. These administrative files can be used as a substitute for a survey or to support surveys (as sampling frames, for imputation, to add new variables, etc.).

Here are some advantages and disadvantages of using administrative data compared to using a census or a sample survey:

Advantages (+)

Lower cost: Using administrative data is less expensive than censuses and surveys because there are no collection operations.

No sampling variance: There is no sampling variability attributed to the statistics produced from administrative data because they are calculated using entire groups of the population.

Time series: Data are collected on an ongoing basis, allowing for trend analysis.

No response burden: Since the data are already collected, there is no additional burden on the respondents.

High level of details: With administrative data, you would be able to produce statistics for small sub-groups of the population or small geographic area, as long as the right classification variables are present in the file and the subgroups have good coverage (i.e. most units in the subgroups are present in the file).

Disadvantages (-)

Less flexibility: Unlike a survey, data users have limited control over which variables are collected. In some case, variables can be limited to a few essential administrative information.

Lower coverage: Data is limited to the population on whom the administrative records are kept. Most of the time, this population is different from the target population which results in sources of under- and over-coverage.

Comparability over time: Definitions are created to serve specific purposes, but often change and evolve over time.

Concepts and definitions: The definitions are established by those who create and manage the file for their own purposes and these definitions might not be fit for use in another context.

Data quality: Data quality can be different from one data provider to another because they don't give the same importance to the different dimensions of quality.

Ethics: With censuses and sample surveys, respondents are aware of what data is being collected. They usually give consent for that data to be used, since the vast majority of surveys are voluntary. With administrative data, it would be difficult to inform and ask consent to all units in the data set. This means individuals and organizations that use administrative data to produce statistical information have a greater responsibility to ensure that data are used in ways that will benefit society and that they have considered data ethics in all steps of the process.

Alternative sources of data

These sources of data are increasingly being used in the production of statistical information to replace or complement traditional methods.

Crowdsourcing involves collecting information from a large community of users and relies on the principle that citizens are the experts of their local environment. Prior to the legalization of cannabis in 2018, the Canadian government needed information on the size and activity of the existing black market for dried cannabis. This information was difficult to gather using a traditional sample survey. For one thing, the characteristic being measured was rare; a probabilistic sample would include many nonusers since they outnumber cannabis users in the Canadian population. Furthermore, some respondents may be hesitant to report the details of their cannabis consumption to an interviewer. Statistics Canada therefore decided to use crowdsourcing to gather the information. The agency established **StatsCannabis**, an anonymous web application that allowed cannabis users to report information about their previous purchases. The Canadian government subsequently used this information to help plan the transition to a legal market for cannabis.

Web scraping is a process through which information is gathered and copied from the web for further analysis. Starting January 2021, web scraped data are being used to model the price of computers and laptops in the

Computer Equipment, Software and Supplies Index of the Consumer Price Index. This change in data collection method aims to improve the coverage of products available on the market and the timeliness of the information on their prices, considering rapid changes to the digital economy. Like for administrative data, users of web scraped data have greater responsibility to consider the ethics of data collection and to follow best practices to avoid inadvertent collection of personal information.

Remote sensing is the acquisition of information about an object or phenomenon from a distant point. Statistics Canada uses remote sensing for its Crop Condition Assessment Program. The growth of vegetation on Canadian farms is observed on a weekly basis using satellite imagery; the data typically become available the same day that the satellite images are processed, allowing near-real-time monitoring of Canadian agriculture. This program provides valuable information while reducing collection costs and easing response burden on producers. Other examples of remote sensing include weather radar systems that track storms, and seismogram arrays that monitor vibrations in the earth.

Statistical registers are data sets created for statistical purposes that are continuously updated with information about all units of a population. They are usually created from the integration of multiple data sources through record linkage and use algorithms or machine learning techniques to consolidate data and derive new variables. Statistics Canada's Business Register is an example of statistical register that is continuously updated from enterprise tax data and sample survey data that is being used as sampling frame for a large number of business surveys and to produce semi-annual Canadian Business Counts.

Finally, **open data** and **big data** are other terms used to describe some types of data. Open data are structured, machine-readable data that are freely shared and that can be used without restrictions. Big data is used to refer to data sets that have such a large number of records and variables that they exceed the capacity of traditional software to process the information with a reasonable time. They are also characterized by the three "v": volume, variety and velocity.

2.3 Exercises

1. List some of the elements (e.g. cost) you would need to consider when choosing a data gathering method.

2. What type of data gathering (census, sample survey or use of administrative data) would be the most appropriate in providing answers to the following questions?

a. What is the main cause of death of young Canadians aged 15 to 25?

b. What type of food should be ordered for a class picnic, based on student preferences?

c. The CEO of a cell-phone company wants to know: If we introduced a new line of services, how would our current clients react?

2.4 Answers

1. When choosing a data gathering method, you should consider the following elements:

- cost (budget);
- time;
- size of population; and
- personnel required to perform the chosen method.

2. What type of data gathering (census, sample survey or use of administrative data) would be the most appropriate in providing answers to the following questions?

a. Administrative data: the information would be provided by death registry.

b. A census, because the number of students in a class is reasonable to survey.

c. A sample survey: it would be more cost-effective and takes less time than a census of every client.

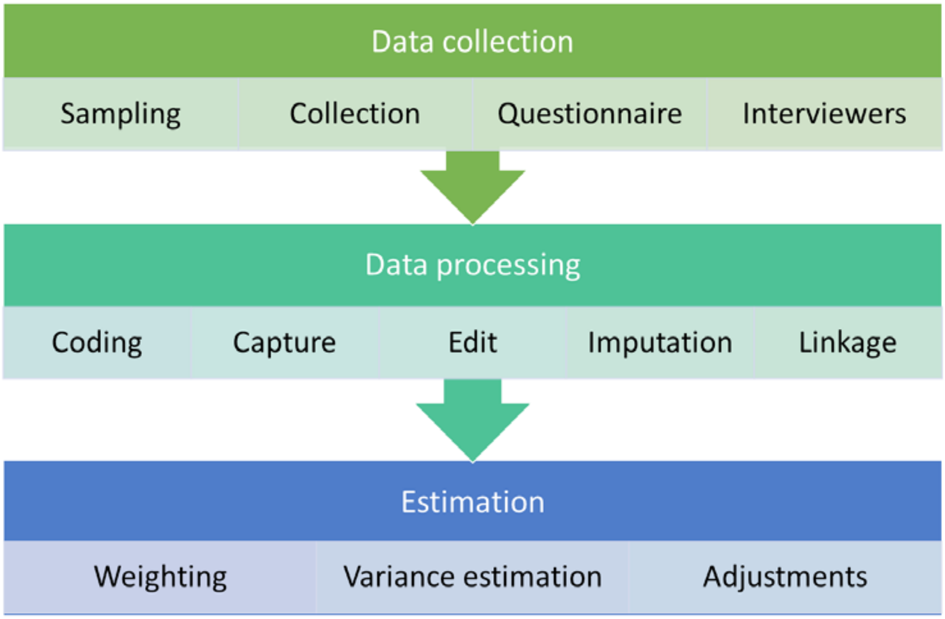
3 Data gathering and processing

Section 2 described who in society is gathering and using data, and for which purposes. Then the advantages and disadvantages of the different types of traditional collection methods were compared. Finally, examples were given of alternative sources of data recently used at Statistics Canada.

Of all these types of data, sample surveys are the most frequently used for producing statistical information about the Canadian society. However, they require careful planning as well as tried and tested methods to select a sample and to carry collection operations. Therefore, the first three subsections of Section 3 are mostly devoted to sample surveys while the rest of the subsections on processing, estimation and quality management can potentially apply to all types of data.

Figure 3.1 summarizes the different steps of data collection, data processing and estimation that will be explored in this section.

Figure 3.1 The different steps of data collection, data processing and estimation



3.1 Planning

At first glance, conducting a survey might appear to be simply asking questions and compiling the answers to obtain statistics. However, it's important to follow precise steps so that the survey results will provide accurate and useful information.

To begin, the following questions should be addressed:

- Why is this survey being conducted?
- Whom will the collected information be about?
- What do I need to know?
- How will the information be used?
- How accurate and timely does the information have to be?

To design a survey, many decisions have to be made about the following issues, which will be covered in detail in the present section.

- Survey objectives
- Target population
- Data requirements
- Choosing the type of collection
- Minimizing error
- Sample size
- Analysis plan
- Questionnaire design
- Data collection methods
- Data processing plan
- Quality control
- Analysis and dissemination of results

Survey objectives

A survey plan begins with objectives that describe why the survey is being done and which population has to be reached to carry this survey. The survey objectives tell a lot about the data that needs to be collected. The objectives also help determine the target population.

Example

Imagine that Ridgemont High School's student council wants to survey students to get information that would help in planning the graduation prom. From this general goal, you can refine objectives. Let's say that the survey objectives are

- To gather information from students in order to determine the factors that will make the prom a success. (The criteria of "success" are that the largest possible number of students will attend the prom and that it will fulfill their expectations.)
- To obtain useful data that will help the prom organizing committee.

The survey plan shows how the objectives will be reached by clearly describing the target population, the data requirements and the variables to be measured, as well as looking at the questions and possible answers and how the data will be processed and analyzed.

Target population

If a survey's objective is to collect information from students, for example, then asking the question "which students?" will help to define the target population.

In the example described previously, the prom organizing committee will probably want to question only students who will be graduating this year, that is, those in the last year of high school (Grade 12). If some of the Grade 12 students are studying part-time and don't intend to graduate this year, they need not be consulted. The target population would therefore be defined as "the full-time Grade 12 graduating students of Ridgemont High School."

Sometimes the target population, that is, the population for which information is required, and the survey population, the population actually being covered by the survey, differ for practical reasons. Ideally, the two populations should be very similar. It is important to note that conclusions based on the survey results will apply only to the survey population.

In our example, some of the full-time Grade 12 graduating students might be away from school at the time of the survey. Since it would be too difficult to reach them, they would not be part of the survey population, although they are part of the target population.

It is also possible that some of the survey concepts and methods that are used may be considered inappropriate for certain segments of the population. For example, consider a survey of post-secondary graduates where the objective is to determine if the graduates found jobs and, if so, what types of jobs. In this case, you might exclude graduates coming from specialized schools such as military schools. These types of graduates would be reasonably assured of securing employment in their field. The target population would therefore be those who graduated from universities, colleges and trade schools.

It may also be necessary to impose geographic limits that will exclude some members of the target population, as some regions may be too difficult or expensive to reach. For example, a business that is doing a survey using in-person interviews may wish to use a sample of the target population living in a densely populated area in order to minimize travel costs.

Data requirements

To determine what kinds of data to collect, ask, “What exactly do we want to know?” and “How will the collected information be used?”

In our example, the organizing committee might consider the following questions:

- Do we need to know the number of students who intend to go to the prom? (This number might also be established from ticket sales.)
- If we ask students whether they intend to go to the prom, should we ask anything in particular to those who don’t intend to go? (By understanding better their reasons for not going, it might be possible to plan certain activities that are of interest thus influencing them to change their minds!)
- When asking about student preferences concerning the prom, what aspects should we consider? Probably elements such as
    - the cost of tickets,
    - the music,
    - the type of refreshments,
    - the day of the week,
    - the venue or location.
- Are there any other factors to consider? Would the students like to have a photographer available? Does everyone want to have a meal before the dance or do some students want just the dance?
- Are students interested in having security guards at the entrance of the venue? What type of transportation would students like to use to get to and from the prom? (The rental of a bus from a central location might be considered.)

When planning a survey, it’s tempting to want to collect as much information as possible. However, the more questions are asked, the longer the survey takes and the more it costs. It’s important to ask: “Do we really need this information?” while considering the time and resources needed to test the questionnaire, process the data and analyze the results.

Another aspect to take into account is the burden the survey imposes on the respondent, so that it's not seen as a nuisance. Response burden is affected by

- the number of questions asked,
- the intrusiveness of the questions,
- the number of times the respondent is contacted (for the same survey or for many surveys),
- the detail of information requested (for example, if asked for a precise income figure, respondents need to consult their official documents, but if asked to choose between five different income ranges, they can answer more easily),
- the time it takes to complete the survey.

Choosing the type of data collection

The level of precision and details pursued and the resources available will determine the choice of the type of data collection. Advantages and disadvantages of different types of collection have been presented on the section on the type of data.

In our example, the organizing committee may decide to do a census of all the graduating students or to survey only a sample of that group.

The type of collection chosen often depends on the **budget** available. Costs are one of the main justifications for choosing to conduct a sample survey instead of a census. With sample surveys, it is possible to obtain valuable results with a relatively small sample of the target population. For example, if you need information on all Canadian citizens over 15 years of age, a survey of a small number of these (1,000 or 2,000 depending on the data requirements) might provide adequate results.

Another advantage of using a sample survey is that it allows investigators to produce information soon after they have identified the need for it, within a rapid turnaround **time**. For example, if an organization wants to measure the public awareness created through an advertising campaign, it should conduct a survey shortly after the campaign is undertaken. Since using a sample of the target population requires a smaller scale of operations, it reduces the data collection and processing time, while allowing more time for planning and quality control.

Minimizing error

When planning a survey, you must be aware of potential sources of error and try to reduce them as much as possible.

In a sample survey, the variation that exists between different samples causes uncertainty called sampling error. For example, let's say you are estimating the average distance between home and school for students in your class of 25 from a sample of 5 persons. Your estimate will depend on which 5 students are sampled. If all 5 sampled students live very close to the school, the results will not be representative of the whole class. It's the variation from one sample to another that causes the sampling error.

As a general rule, the more people surveyed, the smaller the sampling error will be. It is often possible to estimate the sampling error associated with a particular sampling plan, and try to minimize it.

By choosing to do a census, you can avoid errors related to sample variation, but not the other sources of errors, called non-sampling errors. For example, a question might be asked in a way that encourages a certain answer or an error might be made while processing the data or calculating a percentage for a table of results. These types of error must be avoided as much as possible by paying attention to quality control throughout every step of the survey process.

These two types of errors will be discussed in greater details in the section devoted to estimation.

Sample size

Since every sample survey is different, there are no hard and fast rules for determining sample size. The deciding factors are time, cost, operational constraints and the desired precision of the results. Evaluate and assess each of these issues and you will be in a better position to decide the sample size. Also, consider what should be the acceptable level of error in the sample. If there is one characteristic that is central to fulfill the survey objectives and there is a lot of variability of this characteristic in the population, the sample size will need to be bigger to obtain the specified level of precision.

Analysis plan

After identifying all the elements or variables to be measured and preparing the sample design, the next step is the analysis plan—conceiving what the results tables will look like. In other words, you need to plan the tables that you will create for the survey variables. These tables will not yet contain any data, but will show any cross-tabulations you want to make.

In our example, the organizing committee might plan results tables showing the number and percentage for each survey variable (for example, the number and percentage of students who prefer location A to location B for the prom). Some tables could also present cross-tabulations such as “Preferred music by gender.”

These tables help you verify whether the questions you are considering will allow you to reach the survey objectives. They illustrate concretely how the collected information will be used and whether it will adequately measure what you want to know.

Questionnaire design

The questionnaire design is based on the survey’s data requirements and analysis plan. To formulate the questions, it can be helpful to consult the people who will be using the results. You can also consult subject matter experts or look at questions from other surveys on similar topics or themes. Good practices for questionnaire design and testing will be presented in the data collection subsection.

Data collection methods

The choice of the method to gather the required data will have a direct impact on cost, human and material resources, time needed to carry the survey and to assess the quality of data. A first option is the interview, which can be face-to-face or by telephone, with or without computer assistance. The second option is the self-completed questionnaire, which can be paper or online.

Personal interviews are administered by a trained interviewer and can have either a structured or unstructured line of questioning. When done by telephone, questions are structured in a formal interview schedule.

The self-completed questionnaire must be highly structured as the respondent will not have as much help as he would have with an interviewer. It can be returned by mail, through a drop-off system or completed online.

In our example, the organizing committee may opt for a personal interview administered by interviewers who fill out an electronic questionnaire in a spreadsheet program. The interviewer would use a laptop computer to enter the students’ answers into the spreadsheet during the interviews. If some students are concerned about the confidentiality of their answers, the interviewer could give them the option of entering their answers themselves. Such an option, however, might cause more errors and compromise the quality of the collected data, which in turn could increase the time needed for data processing.

Data processing plan

This step deals with transforming the questionnaire responses into output. The tasks involved in data processing include coding, capture, editing, imputation and the creation of derived variables. In short, the aim of this step is to produce a file of data that is free of invalid and missing values and that can be used for estimation and data analysis.

Quality control

This process aims to identify errors and verify results. No matter how much planning and testing goes into a survey, something unexpected will happen. As a result, no survey is ever perfect. Quality control tasks are required to minimize non-sampling errors introduced during various stages of the survey. These tasks include interviewer training, computer program testing, follow-up of non-respondents, and spot-checks of collected responses and output data. Statistical quality-control programs ensure that error levels are kept to a minimum.

Analysis and dissemination of results

After planning data collection and processing, look ahead to the final steps in analyzing and disseminating the results:

- organizing the data using frequency distribution tables;
- summarizing the data using measures of central tendency and measures of dispersion;
- displaying the data through different graph types; and
- writing up the survey’s findings and then disseminating them to the public.

> In our example, members of the prom organizing committee might share the tasks of organizing and analyzing the data, then writing up the conclusions. Decisions about the prom venue, ticket price, type of music, etc. would then be based on these findings. By publishing highlights of the survey in the school newspaper, the student council might demonstrate that its decisions about the prom are based on expectations expressed by the students.

Sections on data exploration and data visualization will present some of these steps with more details.

3.2 Sampling

Many steps of planning a survey rely on a good understanding of information user needs and good knowledge of the steps to follow to carry the survey. One of the first steps for which a good knowledge of mathematical statistics is also needed is sampling, which is the method used to select a subset of units from the target population. There are many methods possible and the chosen method will have a direct impact on the accuracy of the statistics that are being produced. For this reason, it is important to understand the differences between the different sampling designs. They can be split into two types: probability sampling and non-probability sampling. But first, let’s see what elements must be taken into account to choose the sampling design that is the most appropriate in a given context.

3.2.1 Selection of a sample

Sampling allows the estimation of the characteristics of a population by directly observing a portion of the entire population. Researchers are not interested in the sample itself, but in what can be learned from the sample about the entire population. It is essential that a sample survey be correctly defined and organized. If the wrong questions are asked, the collected data will not help in fulfilling the survey objectives. If the questions are asked to the wrong people, the data will not give a good representation of the population. Results will be biased.

Here are the steps to follow to select a sample and to ensure that this sample will allow you to meet the survey’s objective.

Establish the survey's objectives

Specifying the objectives of a survey with as much detail as possible is critical to its ultimate success. The initial users and uses of the data should be identified at this stage. It is also at this stage that a decision should be made on the type of data to be used among census, sample survey, administrative data or an alternative source of data.

Define the target population

No matter which type of data is used, the target population must be well defined. It is the total population for which the information is required. In order to achieve this, the units that make up the population must be described in terms of characteristics that clearly identify them. The following characteristics define the target population:

- Nature of units: persons, hospitals, schools, etc.
- Geographic location: the geographic boundaries of the population have to be determined, as well as the level of geographic detail required for the survey estimate (by province, by city, etc.).
- Reference period: the period covered by the survey.
- Other characteristics, such as socio-demographic characteristics (a particular age group, for example) or type of industry.

Decide on the data to be collected

The data requirements of the survey must be established. It is also necessary to define the terms relative to the data and ensure that these definitions meet data requirements operationally.

Set the level of precision

There is a level of uncertainty associated with estimates from a sample. It is the sampling error. When designing a survey, the acceptable level of uncertainty in the survey estimates has to be established. This level depends on the end use of the results and on the overall budget and time available. The bigger the budget, the more resources available to control quality. The level of uncertainty will also be determined by the sample size. Increasing the sample size will decrease the sampling error. For example, if you sample 24 out of 25 students in your class, there will not be as much sample-to-sample variation as there would be if you only sampled 5 students from among the 25 students in the class.

The sample design

The following steps lead to the determination of the sample design:

1. Determine what the survey population will be (e.g. students, men aged 20 to 35, newborn babies, etc.).
2. Choose the most appropriate survey time frame.
3. Define the survey units.
4. Establish the sample size (e.g. a sample of 100 from a population of 1,000).
5. Select a sampling method.

Estimation techniques to be used, that is, how results will be generalized to the entire population and how sampling error will be calculated, will result directly from the sampling design and will be discussed in the upcoming section on estimation.

The survey population

Some members of the target population have to be excluded because of operational constraints such as the high cost of collecting data in some remote areas, the difficulty of identifying and contacting certain components of the target population, etc. The population that is effectively included is called the survey population or the **observed population**. The target population is the population we **want to observe** while the survey population is the population we **can observe**.

The goal is to have the survey population as close as possible to the target population. It is also very important to inform the users of the data of the differences between the two populations, as the results of the survey will apply only to the survey population.

For example, a target population for a survey could be all Canadians aged 15 years and over (on a particular reference date), while the survey population could exclude residents of the Yukon, Nunavut and the Northwest Territories, persons living on Aboriginal reserves, full-time members of the Canadian Armed Forces and residents of institutions. These Canadians might be excluded for various reasons: to survey people in the territories might prove to be difficult and expensive, military personnel may not be available for surveying if they are out on a mission, etc. Using this example, about 2% of the target population would be excluded from the survey population.

The survey frame

The survey frame, also called the sampling frame, is the tool used to gain access to the population. There are two types of frames: list frames and area frames. A list frame is just a list of the units in a population. Each unit can be identified and the frame includes the information needed to access these units. A good frame should be complete and up-to-date. No member of the survey population should be excluded from the frame or appear more than once and no unit that is not part of the population (e.g. deceased persons) should be on the frame. The chosen frame will impact the selected survey population. For instance, if a list of telephone numbers is used to select a sample of households, then all households without telephones are excluded from the survey population.

An area frame is a list of geographic areas. Instead of selecting units directly from the frame as one would with a list frame, geographic areas are selected and a means to access units located in these areas is identified, like visiting the units in person, for instance. Suppose that you were surveying a rural town in Quebec to see what percentage of residents are farmers. If you were provided with an area frame, then you would be able to locate which roads to visit, but you would still have to find out the names and addresses of the residents on each road.

The survey units

There are three types of units that have to be accurately identified in order to avoid problems during the selection, data collection and data analysis stages. They are as follows:

- The sampling unit is part of the frame and therefore subject to being selected.
- The respondent unit, or reporting unit, who provides the information needed by the survey.
- The unit of reference, or unit of analysis, the unit about which information is provided and who is used to analyze the survey results.

For example, in a survey about newborns in Edmonton, the sampling unit might be a household, the reporting unit one of the parents or a legal guardian, and the unit of reference the baby.

The sampling units may differ depending on the survey frame used. This is why the survey population, survey frame and survey units are defined in conjunction with one another.

The sample size

The level of precision needed for the survey estimates will impact the sample size. However, it is not as easy to determine the sample size as one may think. Generally, the actual sample size of a survey is a compromise between the level of precision to be achieved, the survey budget and any other operational constraints. In order to achieve a certain level of precision, the sample size will depend, among other things, on the following factors:

- The variability of the characteristics being observed: If every person in a population had the same salary, then a sample of one person would be all you need to estimate the average salary of the population. If the salaries are very different, then you would need a bigger sample in order to produce a reliable estimate.
- The population size: To a certain extent, the bigger the population, the bigger the sample needed. But once you reach a certain size, an increase in population no longer affects the sample size. For instance, the necessary sample size to achieve a certain level of precision will be about the same for a population of one million as for a population twice that size.

- The sampling and estimation methods: Not all sampling and estimation methods have the same level of efficiency. The more efficient the method, the smaller the sample size needed to obtain a given precision of the estimates. However, because of operational constraints and limitations of the survey frames, you cannot always use the most efficient method.

The sampling method

There are two types of sampling methods: probability sampling and non-probability sampling. The difference between them is that in probability sampling, every unit has a probability of being selected that can be quantified. This is not true for non-probability sampling. The next section will describe features of both types of sampling and detail some of the methods related to each type.

3.2.2 Probability sampling

Probability sampling refers to the selection of a sample from a population, when this selection is based on the principle of randomization, that is, random selection or chance. Probability sampling is more complex, more time-consuming and usually more costly than non-probability sampling. However, because units from the population are randomly selected and each unit's selection probability can be calculated, reliable estimates can be produced and statistical inferences can be made about the population.

There are several different ways in which a probability sample can be selected.

When choosing a probability sample design, the goal is to minimize the sampling error of the estimates for the most important survey variables, while simultaneously minimizing the time and cost of conducting the survey. Some operational constraints can also have an impact on that choice, such as characteristics of the survey frame.

In the present section, each of these methods will be described briefly and illustrated with examples.

Simple random sampling

In **simple random sampling (SRS)**, each sampling unit of a population has an equal chance of being included in the sample. Consequently, each possible sample also has an equal chance of being selected. To select a simple random sample, you need to list all of the units in the survey population.

Example 1

To draw a simple random sample from a telephone book, each entry would need to be numbered sequentially. If there were 10,000 entries in the telephone book and if the sample size was 2,000, then 2,000 numbers between 1 and 10,000 would need to be randomly generated by a computer. All numbers would have the same chance of being generated by the computer. The 2,000 telephone entries corresponding to the 2,000 computer-generated random numbers would make up the sample.

SRS can be done with or without replacement. An SRS with replacement means that there is a possibility that the sampled telephone entry may be selected twice or more. Usually, the SRS approach is conducted without replacement because it is more convenient and gives more precise results. In the rest of the text, SRS will be used to refer to SRS without replacement, unless stated otherwise.

SRS is the most commonly used method. The advantage of this technique is that it does not require any information on the survey frame other than the complete list of units of the survey population along with contact information. Also, since SRS is a simple method and the theory behind it is well established, standard formulas exist to determine the sample size, the estimates and so on, and these formulas are easy to use.

On the other hand, this technique necessitates a list of all units of the population. If such a list doesn't already exist and the target population is large, it can be very expensive or unrealistic to create one. If a list already exists and includes auxiliary information on the units, then the SRS is not taking advantage of information that allows other methods to be more efficient (like stratified sampling, for example). If collection has to be made in-person, SRS could give a sample that is too spread out across multiple regions, which could increase costs and duration of the survey.

Example 2

Imagine that you own a movie theatre and you are offering a special horror movie film festival next month. To decide which horror movies to show, you survey moviegoers to ask them which of the listed movies are their favorites. To create the list of movies needed for your survey, you decide to sample 10 of the 100 best horror movies of all time. One way of selecting a sample would be to write all of the movie titles on slips of paper and place them in an empty box. Then, draw out 10 titles and you will have your sample. By using this approach, you will have ensured that each movie had an equal probability of selection. You could even calculate this probability of selection by dividing the sample size (n=10) by the population size of the 100 best horror movies of all time (N=100). This probability would be 0.10 (10/100) or 1 in 10.

Systematic sampling

Systematic sampling means that there is a gap, or interval, between each selected unit in the sample. For instance, you could follow these steps:

1. Number the units on your frame from 1 to **N** (where **N** is the total population size).
2. Determine the sampling interval (**K**) by dividing the number of units in the population by the desired sample size. For example, to select a sample of 100 from a population of 400, you would need a sampling interval of 400/100 = 4. Therefore, **K** = 4. You will need to select one unit out of every four units to end up with a total of 100 units in your sample.
3. Select a number between one and **K** at random. This number is called **the random start** and it would be the first number included in your sample. If you choose 3, the third unit on your frame would be the first unit included in your sample; if you choose 2, your sample would start with the second unit on your frame.
4. Select every **Kth** (in this case, every fourth) unit after that first number. For example, the sample might consist of the following units to make up a sample of 100: 3 (the random start), 7, 11, 15, 19 …395, 399 (up to **N**, which is 400 in this case).

In the example above, you can see that there are only four possible samples that can be selected, corresponding to the four possible random starts:

1, 5, 9, 13 … 393, 397

2, 6, 10, 14 … 394, 398

3, 7, 11, 15 … 395, 399

4, 8, 12, 16 … 396, 400

Each member of the population belongs to only one of the four samples and each sample has the same chance of being selected. From that, we can see that each unit has a one in four chance of being selected in the sample. This is the same probability as if a simple random sample of 100 units was selected. The main difference is that with SRS, any combination of 100 units would have a chance of making up the sample, while with systematic sampling, there are only four possible samples. The units' order on the frame will determine the possible samples for systematic sampling. If the population is randomly distributed on the frame, then systematic sampling should yield results that are similar to simple random sampling.

This method is often used in industry, where an item is selected for testing from a production line to ensure that machines and equipment are of a standard quality. For example, a tester in a manufacturing plant might perform a quality check on every 20th product in an assembly line. The tester might choose a random start between the numbers 1 and 20. This will determine the first product to be tested; every 20th product will be tested thereafter.

Interviewers can use this sampling technique when questioning people for a sample survey. The market researcher might select, for example, every 10th person who enters a particular store, after selecting the first person at random. The surveyor may interview the occupants of every fifth house on a street, after randomly selecting one of the first five houses.

The advantages of systematic sampling are that the sample selection cannot be easier: you only get one random number, the random start, and the rest of the sample automatically follows. The biggest drawback of the systematic sampling method is that if there is some periodical feature in the way the population is arranged on a list and that periodical feature coincides in some way with the sampling interval, the possible samples may not be representative of the population. This can be seen in the following example:

Example 3

Suppose you run a large grocery store and have a list of the employees in each section. The grocery store is divided into the following 10 sections: deli counter, bakery, cashiers, stock, meat counter, produce, pharmacy, photo shop, flower shop and dry cleaning. Each section has 10 employees, including a manager (making 100 employees in total). Your list is ordered by section, with the manager listed first and then, the other employees by descending order of seniority.

If you wanted to survey your employees about their thoughts on their work environment, you might choose a small sample to answer your questions. If you use a systematic sampling approach and your sampling interval is 10, then you could end up selecting only managers or only the newest employees in each section. This type of sample would not give you a complete or appropriate picture of your employees' thoughts.

Sampling with probability proportional to size

Probability sampling requires that each member of the survey population has a known probability of being included in the sample, but it does not require that this probability be the same for everyone. If there is information available on the frame about the size of each unit (e.g. number of employees for each business) and if those units vary in size, this information can be used in the sampling selection in order to increase the efficiency. This is known as **sampling with probability proportional to size** (PPS). With this method, the bigger the size of the unit, the higher the chance of being included in the sample. For this method to bring increased efficiency, the measure of size needs to be accurate. This is a more complex sampling method that will not be discussed in further detail here.

Stratified sampling

When using **stratified sampling**, the population is divided into homogeneous, mutually exclusive groups called strata, and then independent samples are selected from each stratum. Any of the sampling methods mentioned in this section can be used to sample within each stratum. The sampling method can vary from one stratum to another. A population can be stratified by any variable for which a value is available for all units on the sampling frame prior to sampling (e.g. age, sex, province of residence, income).

Why create strata? There are many reasons, the main one being that it can make the sampling strategy more efficient. It was mentioned in the previous section that in order to an estimation of a certain precision, a larger sample size is needed for a characteristic that varies greatly from one unit to the other than for a characteristic with smaller variability. For example, if every person in a population had the same salary, then a sample of one individual would be enough to get a precise estimate of the average salary.

This is the idea behind the efficiency gain obtained with stratification. If you create strata within which units share similar characteristics and are considerably different from units in other strata then you would only need a small sample from each stratum to get a precise estimate of total income for that stratum. Then you could combine these estimates to get a precise estimate of total income for the whole population. If you were to use a SRS in the whole population without stratification, the sample would need to be larger than the total of all stratum samples sizes to get an estimate of total income with the same level of precision.

Another advantage is that stratified sampling ensures an adequate sample size for subgroups of interest in the population. When a population is stratified, each stratum becomes an independent population and a sample size is calculated for each of them.

Example 4

Suppose you want to estimate how many high school students have part-time jobs at the national level and provincial level. If you were to select a simple random sample of 25,000 people from a list of all high school students in Canada (assuming such a list was available for selection), you would end up with just a little over 100 people from Prince Edward Island, since they account for less than 0.5% of the Canadian population. This sample would probably not be large enough for the kind of detailed analysis you were planning for. Stratifying your list by province and then determining a sample size needed in each province would allow you to get the required level of precision for Prince Edward Island and for each of the other provinces as well.

Stratification is most useful when the stratifying variables are

- simple to work with,
- easy to observe,
- closely related to the topic of the survey.

Cluster sampling

Sometimes it is too expensive to have a sample too spread out geographically. Travel costs can become expensive if interviewers have to survey people from one end of the country to the other. To reduce costs, statisticians may choose a **cluster sampling** technique.

Cluster sampling divides the population into groups or clusters. A number of clusters are selected randomly to represent the total population, and then all units within selected clusters are included in the sample. No units from non-selected clusters are included in the sample. They are represented by those from selected clusters. This differs from stratified sampling, where some units are selected from each stratum. Examples of clusters are factories, schools and geographic areas such as electoral subdivisions.

Example 5

Suppose you are a representative from an athletic organization wishing to find out which sports Grade 11 (or secondary 4) students are participating in across Canada. It would be too costly and lengthy to survey every Canadian in Grade 11, or even a couple of students from every Grade 11 class in Canada. Instead, 100 schools are randomly selected from all over Canada. These 100 schools are the sampled clusters. Then all Grade 11 students in all 100 clusters are surveyed.

Cluster sample creates “pockets” of sampled units instead of spreading the sample over the whole territory, which allows for cost reduction in collection operations. Another reason to use cluster sampling is that sometimes a list of all units in the population is not available, while a list of all clusters is either available or easy to create.

In most cases, cluster sampling is less efficient than SRS. This is the main drawback of the method. For this reason, it is usually better to survey a large number of small clusters instead of a small number of large clusters. Why? Because neighbouring units tend to be more alike, resulting in a sample that does not represent the whole spectrum of opinions or situations present in the overall population. In the example 5, students in the same school tend to participate in the same types of sports, that is, those for which the facilities are available at their school.

Another drawback to cluster sampling is that you do not have total control over the final sample size. Since not all schools have the same number of Grade 11 students and you must interview every student in your sample, the final size may be larger or smaller than you expected.

Multi-stage sampling

Multi-stage sampling is like cluster sampling, except that it involves selecting a sample within each selected cluster, rather than including all units from the selected clusters. This type of sampling requires at least two stages. In the first stage, large clusters are identified and selected. In the second stage, units are selected from within the selected clusters using any of the probability sampling methods. In this context, the clusters are referred to as primary sampling units (PSU) and units within clusters are referred to as secondary sampling units (SSU). When there are more than two stages, tertiary sampling units (TSU) are selected within SSU, and the process continues until there is a final sample.

Example 6

In Example 5, a cluster sample would choose 100 schools and then interview every Grade 11 student from those schools. Instead, you could select more schools, get a list of all Grade 11 students from these selected schools and select a random sample of Grade 11 students from each school. This would be a two-stage sampling design. Schools would be the PSU and students the SSU.

You could also get a list of all Grade 11 classes in the selected schools, pick a random sample of classes from each of those schools, get a list of all the students in the selected classes and finally select a random sample of students from each selected class. This would be a three-stage sampling design. Schools would be the PSU, classes would be the SSU and students would be the TSU. Each time a stage is added, the process becomes more complex.

Now imagine that each school has on average 80 Grade 11 students. Cluster sampling would then give your organization a sample of about 8,000 students (100 schools x 80 students). If you wanted a bigger sample, you could select schools with more students. For a smaller sample, you could select schools with fewer students. One way to control the sample size would be to stratify the schools into large, medium and small sizes (in terms of the number of Grade 11 students) and select a sample of schools from each stratum. This is called **stratified cluster sampling**.

As an alternative method, you could use a three-stage design. You would select a sample of 400 schools, then select two Grade 11 classes per school and finally, select 10 students per class. This way, you still end up with a sample of about 8,000 students (400 schools x 2 classes x 10 students), but the sample would be more spread out.

You can see from this example that with multistage sampling, you still have the benefit of a more concentrated sample for cost reduction. However, the sample is not as concentrated as cluster sampling and the sample size needed to obtain a given level of precision would still be bigger than for an SRS because the method is less efficient. Nonetheless, multistage sampling could still save a large amount of time and effort compared to SRS because you would not need to have a list of all Grade 11 students. All you would need is a list of the classes from the 400 schools and a list of the students from the 800 classes.

Multi-phase sampling

A **multi-phase sample** collects basic information from a large sample of units and then collects more detailed information for a subsample of these units. The most common form of multi-phase sampling is two-phase sampling (or double sampling), but three or more phases are also possible.

Multi-phase sampling is quite different from multistage sampling, despite the similarity of their names. Although multi-phase sampling also involves taking two or more samples, all samples are drawn from the same frame. Selection of a unit in the second phase is conditional to its selection in the first phase. A unit not selected in the first phase will not be part of the second-phase sample. Like for multistage sampling, the more phases used, the more complex the sample design and estimation.

Multi-phase sampling is useful when the sampling frame lacks auxiliary information that could be used to stratify the population or to screen out part of the population.

Example 7

Suppose that an organization needs information about cattle farmers in Alberta, but the survey frame lists all types of farms—cattle, dairy, grain, hog, poultry and produce. To complicate matters, the survey frame does not provide any auxiliary information for the farms listed there.

A simple survey whose only question would be "Is part or all of your farm devoted to cattle farming?" could be conducted. With only one question, this survey should have a low cost per interview (especially if done by telephone) and, consequently, the organization should be able to draw a large sample. Once the first sample has been drawn, a second, smaller sample can be extracted from among the cattle farmers and more detailed questions asked of these farmers. Using this method, the organization avoids the expense of surveying units that are not in this specific scope (i.e. non-cattle farmers).

In the example 7, data collected in the first phase has been used to exclude units that are not part of the target population. In another context, this data could have been used to improve the efficiency of the second phase, by creating strata, for example. Multi-phase sampling can also be used to reduce response burden or when there are very different costs associated with different questions of a survey, as illustrated in the next example.

Example 8

In a health survey, participants are asked some basic questions about their diet, smoking habits, exercise routines and alcohol consumption. In addition, the survey requires that respondents subject themselves to some direct physical tests, such as running on a treadmill or having their blood pressure and cholesterol levels measured.

Filling out questionnaires or interviewing participants are relatively inexpensive procedures, but the medical tests require the supervision and assistance of a trained health practitioner, as well as the use of an equipped laboratory, both of which can be quite costly. The best way to conduct this survey would be to use a two-phase sample approach. In the first phase, the interviews are performed on an appropriately sized sample. From this sample, a smaller sample is drawn. Only participants selected in the second sample would take part in the medical tests.

3.2.3 Non-probability sampling

Non-probability sampling is a method of selecting units from a population using a subjective (i.e. non-random) method. Since non-probability sampling does not require a complete survey frame, it is a fast, easy and inexpensive way of obtaining data. However, in order to draw conclusions about the population from the sample, it must assume that the sample is representative of the population. This is often a risky assumption to make in the case of non-probability sampling due to the difficulty of assessing whether the assumption holds. In addition, since elements are chosen arbitrarily, there is no way to estimate the probability of any one element being included in the sample. Also, no assurance is given that each item has a chance of being included, making it impossible either to estimate sampling variability or to identify possible bias.

In general, official statistical agencies around the world have been using probability sampling as their preferred tool to meet information needs about a population of interest. In the last few years, however, there have been some research and studies about how to apply non-probability sampling into the official statistics. Using other data sources has been increasingly explored. There are five key reasons behind this trend:

- the decline in response rates in probability surveys;
- the high cost of data collection;
- the increased burden on respondents;
- the desire for access to real-time statistics, and
- the surge of non-probability data sources such as web surveys and social media.

Some have suggested the possibility of a shift in the paradigm and traditional approach to statistics. However, data from non-probability sources have a few challenges with respect to data quality, including the potential presence of participation and selection bias. Therefore, data collected using non-probability sampling should be used with extra caution.

The commonly used non-probability sampling methods include the following.

Convenience or haphazard sampling

Units are selected in an arbitrary manner with little or no planning involved. Haphazard sampling assumes that the population units are all alike, then any unit may be chosen for the sample. An example of haphazard sampling is the vox pop survey where the interviewer selects any person who happens to walk by. Unfortunately, unless the population units are truly similar, selection is subject to the biases of the interviewer and whoever happened to walk by at the time of sampling.

Volunteer sampling

The respondents are only volunteers in this method. Generally, volunteers must be screened so as to get a set of characteristics suitable for the purposes of the survey (e.g. individuals with a particular disease). This method can be subject to large selection biases, but is sometimes necessary. For example, for ethical reasons, volunteers with particular medical conditions may have to be solicited for some medical experiments.

Another example of volunteer sampling is callers to a radio or television show, when an issue is discussed and listeners are invited to call in to express their opinions. Only the people who care strongly enough about the subject one way or another tend to respond. The silent majority does not typically respond, resulting in a large selection bias. Volunteer sampling is often used to select individuals for focus groups or in-depth interviews (i.e. for qualitative testing, where no attempt is made to generalize to the whole population).

Judgement sampling

With this method, sampling is done based on previous ideas of population composition and behaviour. An expert with knowledge of the population decides which units in the population should be sampled. In other words, the expert purposely selects what is considered to be a representative sample. Judgment sampling is subject to the researcher's biases and is perhaps even more biased than haphazard sampling.

Since any preconceptions the researcher has are reflected in the sample, large biases can be introduced if these preconceptions are inaccurate. However, it can be useful in exploratory studies, for example in selecting members for focus groups or in-depth interviews to test specific aspects of a questionnaire.

Quota sampling

This is one of the most common forms of non-probability sampling. Sampling is done until a specific number of units (quotas) for various subpopulations have been selected. Quota sampling is a means for satisfying sample size objectives for the subpopulations.

The quotas may be based on population proportions. For example, if there are 100 men and 100 women in the population and a sample of 20 are to be drawn, 10 men and 10 women may be interviewed. Quota sampling can be considered preferable to other forms of non-probability sampling (e.g. judgment sampling) because it forces the inclusion of members of different subpopulations.

Quota sampling is somewhat similar to stratified sampling, which is probability sampling, in that similar units are grouped together. However, it differs in how the units are selected. In probability sampling, the units are selected randomly while in quota sampling a non-random method is used—it is usually left up to the interviewer to decide who is sampled. Contacted units that are unwilling to participate are simply replaced by units that are, in effect ignoring nonresponse bias. Market researchers often use quota sampling (particularly for telephone surveys) instead of stratified sampling to survey individuals with particular socio-economic profiles. This is because compared with stratified sampling, quota sampling is relatively inexpensive and easy to administer and has the desirable property of satisfying population proportions. However, it disguises potentially significant selection bias.

As with all other non-probability sample designs, in order to make inferences about the population, it is necessary to assume that persons selected are similar to those not selected. Such strong assumptions are rarely valid.

Snowball or network sampling

Suppose a researcher wishes to find rare individuals in the population, and already knows of the existence of some of these individuals and how to contact them. One approach is to contact those individuals and simply ask them if they know anyone like themselves, then contact those people, etc. The sample grows like a snowball rolling down a hill to hopefully include virtually everybody with that characteristic. Snowball sampling is useful for rare or hard to reach populations such as people with disabilities, homeless people, drug users, or other persons who may not belong to an organised group or such as musicians, painters, or poets, not readily identified on a survey list frame. However, some individuals or subgroups may have no chance of being sampled. In order to be able to generalize the conclusion to the whole population, some assumptions, which are usually not met, are required.

Crowdsourcing

Crowdsourcing has been defined slightly differently by researchers from various areas. Despite the multiplicity of definitions for crowdsourcing, one constant has been the broadcasting of a problem to the public, and an open call for contributions to help solve the problem. Members of the public submit solutions that are then owned by the entity (e.g. individuals, companies, or organizations), which originally broadcast the problem. Crowdsourcing is channelling the experts' desire to solve a problem and then freely sharing the answer with everyone.

As part of Statistics Canada's modernization, crowdsourcing has become an innovative way to collect valuable information for statistical purposes. By using crowdsourcing as the only collection method, surveys can be executed quickly with reduced cost and response burden. To better understand the challenges associated with crowdsourcing and to ensure that the results are good quality, methods are being developed to compare and validate the data with other sources of complementary data. A couple of examples are outlined below.

- As part of the OpenStreetMap (OSM) pilot project, which was completed in March 2018, crowdsourced geographic information was collected by mapping the building footprints in the Ottawa, Ontario and Gatineau, Quebec areas. The network and experience of this pilot project helped to launch the Building Canada 2020 initiative (BC2020), aimed at mapping all building footprints of Canada on OSM by the year 2020.

Web panels

A web panel (or online or internet panel) could be defined as an access panel of people willing to respond to web questionnaires. It contains a sample of potential respondents who declare that they will cooperate for future data collection if selected. A web panel survey is a survey utilizing samples from web panels.

Web panels can be seen as sampling frames for web panel surveys. All persons in the panels must have up-to-date e-mail addresses. Recruitment for web panels can be made in different ways. Respondents can be sourced from offline channels: telephone, TV ads, radio ads, ads in newspapers and magazines, addressed letters, outdoor posters, customer registers, etc. Respondents can also be sourced from online channels: e-mails, websites, banners, community sites, member programs, etc. Often, many channels are used in order to achieve the necessary diversity. After the recruitment, a profile survey is conducted in order to collect information on the new participants to the panel. The recruitment can be done using either probability-based or self-recruited panels. In practice, the distinction between these two may not be very important if the nonresponse rate is very high for the probability-based panels. Sometimes incentives, such as gift cards or souvenirs, are used to attract people and boost response rates. Web panels are often used for marketing research or pilot studies.

During the pandemic of COVID-19, Statistics Canada developed a new web panel survey, Canadian Perspectives Survey Series (CPSS), to get timely information about how Canadians are coping with COVID-19. More than 4,600 people in the 10 provinces responded to this survey between March 29 and April 3. Unlike the most web panels, CPSS is a probabilistic panel based on the Labour Force Survey (LFS), as some respondents agreed to complete short online questionnaires following their participation to the LFS. CPSS enables Statistics Canada to collect important information from Canadians more efficiently, more rapidly and at a lower cost, compared with traditional survey methods.

Advantages and disadvantages of non-probability sampling

Advantages (+)

Quick and convenient

As a general rule, non-probability samples can be constituted quickly, which allows the survey to be launched, executed and finished in shorter times.

Inexpensive

It usually only takes a few hours to an interviewer to conduct such a survey. As well, non-probability samples are generally not spread out geographically, therefore travelling expenses for interviewers are low. In web panels or crowdsourcing, no interviewers are necessary. Tracing and persuasion of non-respondents are not required or less demanding.

Reduce respondent burden

In the case of volunteer sampling or crowdsourcing, respondents volunteer to participate in the survey without being solicited personally.

Disadvantages (-)

Selection bias

In order to make inferences about the population, it requires strong assumptions about the similarity between the sample and the population even though the respondents are self-selected. Due to the selection bias presented in all non-probability samples, these are often dangerous assumptions to make. When generalization to the whole population is to be made, probability sampling should be performed instead.

Noncoverage (undercoverage) bias

Since some units in the population can have no chance of being included in the sample, it results noncoverage bias. For example, people without the internet at home might never be selected for a web panel and may differ from those with the internet.

Difficulty of assessing the quality

It is impossible to determine the probability that a unit in the population is selected for the sample, so reliable estimates and estimates of sampling error cannot be computed.

3.3 Collecting

In the section on sampling, we presented the different methods to select a subset of the units of a population to carry a sample survey. In this section, we will present the different methods to collect data from this sample. The three important parts of collection are the choice of a collection mode, questionnaire design and the role of interviewers. They are as important in the survey as the sampling plan. First because of their impact on the budget and duration of the survey, but mostly because of their impact on data quality. The right collection strategy should aim to decrease nonresponse as much as possible while a good questionnaire should be developed with the goal of minimizing the measurement error. Nonresponse and measurement error are the two main sources on non-sampling error.

3.3.1 Data collection methods

The main modes of data collection for surveys are in-person interviews, telephone interviews and self-completed questionnaires. Interviews can be computer-assisted or not, and self-completed questionnaires can be on paper or digital support. To choose the best method, factors such as the characteristics of the sample frame, the target population, the survey budget, the desired level of accuracy, the sensitivity of the information to be collected or the complexity of the survey concepts must be taken into account. Here is an overview of the main collection modes.

In-person interview

Interviewers visit people in their environment to survey them. It's a good way to get higher response rates to a sample survey or census and, because of the interviewers' work, the quality of collected data is higher. However, travel costs can be high. When the interview takes place at the place of residence and nobody in the household is present or available to reply to the survey, the interviewer might have to come back multiple times.

When the interview is computer-assisted, the interviewer brings a laptop or electronic tablet and capture the data directly in a database or with the help of an electronic questionnaire designed for this purpose. This method reduces the time needed for data processing and the interviewer avoid carrying a large amount of paper questionnaires which have to be stored safely to protect confidentiality of respondents. However, it is more expensive and longer to implement because computer hardware must be provided to interviewers and the computer systems need to be developed and tested before the start of data collection. When the interview is not computer-assisted, the interviewer writes down the answers on a paper questionnaire. This method requires fewer resources and preparation before collection, but increases data processing time since the answers must be captured in a digital support to be processed and analyzed with software.

Telephone interview

Interviewers call people to survey them. This collection mode is faster and less expensive than an in-person interview. But you need a sample frame with the phone numbers of the population units and the observed population exclude households without a telephone. It is also very easy for the person called to hang up the phone … or simply, not to answer the phone in the first place if they don't know the incoming telephone number. For this reason, response rates are lower than in-person interviews. It is also important to note that fewer and fewer households have a landline nowadays which is likely to make it more difficult to reach them in the future.

Like for an in-person interview, the telephone interview can be computer-assisted or not. Computer-assisted means more preparation before collection, but less processing data after data collection.

Self-completed questionnaire

A questionnaire is provided to the respondent who has to fill and return it. It can be printed or it can be a hyperlink and access code to fill online. In both cases, it can be handed directly, sent by mail or delivered at the dwelling. To send the questionnaire by mail you need a list of addresses in the sampling frame. Home delivery can be useful when dwellings are not listed in the sampling frame.

It's the least expensive mode, and it can be used to reach a very large number of people. The respondent can fill the survey at the moment of his choice. Self-completed questionnaires may be easier when the survey questions are sensitive. The drawbacks of this mode are that response rates are lower than for other collection mode and the quality of collected data is not as good. It is also the slowest method because there is no control over the moment at which the respondent will fill and return the questionnaire.

The questionnaire must be simple with very precise instruction because the respondent cannot ask for clarification if they don't understand the question. Respondent with limited capacities in reading or writing in French and English can have difficulties in filling the questionnaire.

When a digital support is used, often called Electronic Questionnaire (EQ), the web interface can be more convivial for the respondent and allow him to answer the survey faster. It is possible to notify the respondent that a question has not been answered and to indicate to him his progression in the questionnaire. Return of the questionnaire is almost instantaneous, and the respondent doesn't have to remind himself to return it by mail. Like for computer-assisted interviews, processing time is reduced because the answers are captured directly in a machine-readable format. However, people with no access to the internet or with less computer knowledge will be less likely to participate unless they are given the option to answer on a printed form.

Nowadays, EQ is used as the first option before telephone interviews by many surveys, including business and household surveys.

Other methods

Other methods are direct observation or direct measurement. Here are two examples. In the context of data collection for the Consumer Price Index, surveyors visit points of sale to note the prices of a predetermined list of items. This data allows the estimation of the inflation rate in Canada. In the Canadian Health Measures Survey, some physical measures are taken from respondents like weight, blood pressure and blood measures. These data are gathered using mobile examination centres or mobile clinics.

Combining methods

The most efficient strategy will often use a combination of methods that were just described. The Canadian Census of the population is a good example of a multimode collection strategy. During Census 2016, a letter was sent by mail to a majority of households, with a hyperlink and a secured access code to fill the census form online, along with a free phone number to request a printed form. In areas where not all dwellings are listed on the census frame, printed forms were delivered by census enumerators responsible for the listing operations. Along with the print form, a hyperlink and a secured access code to fill online were also provided. In Indigenous reserves and hard-to-reach regions, like Nunavut for example, data collection was carried with in-person interviews. Finally, in-person interviews and telephone interviews were used for nonresponse follow-up. The similar strategy has been applied to Census 2021.

A communication strategy used along with the data collection strategy can be useful to make the survey known to the public and increase response rates. It can be a letter sent by mail to initiate contact with the household, a press release in local media or, more rarely, a national communication strategy like in the case of the census of the population.

What is paradata?

Paradata is data collected during the collection operations, but which is about the collection itself and not about the survey subject. It can be the time at which the interview was done, the length of the interview, the number of times it took to reach the respondents, etc. It can be useful in the conception of adaptive survey designs or as auxiliary variables to adjust results for nonresponse.

3.3.2 Questionnaire design

Questionnaires play a central role in the data collection process. A well-designed questionnaire efficiently collects the required data with a minimum number of errors. It facilitates the coding and capture of data and it leads to an overall reduction in the cost and time associated with data collection and processing. The biggest challenge in developing a questionnaire is to translate the objectives of the survey into a well conceptualized and methodologically sound study.

Before you can design the questionnaire, you must plan the survey as a whole, including the objectives, data needs and analysis. Once the questionnaire is designed, it must be tested before you can proceed with the data collection.

There is a lot to consider when developing a questionnaire. The following is a list of some key points to think about:

- Is the introduction informative? Does it stimulate respondent interest?
- Are the words simple, direct and familiar to all respondents?
- Do the questions read well? Does the overall questionnaire flow well?
- Are the questions clear and as specific as possible?
- Does the questionnaire begin with easy and interesting questions?
- Is there a specific time reference?
- Are any of the questions double-barrelled?
- Are any questions leading or loaded?
- Should the questions be open-ended or close-ended? If the questions are close-ended, are the response categories mutually exclusive and collectively exhaustive?
- Are the questions applicable to all respondents?

Introduction and conclusion of the questionnaire

The introduction of the questionnaire is very important because it outlines the pertinent information about the survey. The introduction should:

- provide the title or subject of the survey;
- identify the sponsor;
- explain the purpose of the survey;
- request the respondent's co-operation;
- inform the respondent about confidentiality issues;
- specify whether the survey is (voluntary or mandatory), and
- state whether there are any data-sharing agreements with other organizations

Respondents frequently question the value of the gathered information to themselves and to others. Therefore, be sure to explain why it is important to complete the questionnaire, how the information will be used, and how respondents can access the results. Ensuring that respondents understand the value of their information is vital in undertaking a survey.

The following is an example of a good introduction to a questionnaire.

Assessing Student Needs

School name________________________________

Please take some time (approximately 50 to 75 minutes) to complete this questionnaire. Your responses will provide important information that will help your school in planning better ways to support your health and well-being.

Confidential

What this survey is for?

This survey provides you with an opportunity to share your thoughts on what is needed to keep you and your school safe and healthy.

You do not have to complete this survey if you do not wish to do so. However, everyone's views are important and the more participation we receive, the better the results will be. Please understand that this questionnaire is completely confidential.

1. **Do not** write your name on the questionnaire.
2. Seal your questionnaire in the envelope provided.

Once the envelope is sealed, it will *only* be opened by the team entering your responses to the questions into the computer system. Your envelope will be placed with many others and there will be no way to identify individual respondents. The results of ***all*** the questionnaires will be added together and reported back to the school.

The opening questions of any survey should establish the respondents' confidence in their ability to answer the remaining questions. If necessary, the opening questions should help determine whether the respondent is a member of the survey population.

A good questionnaire ends with a comments section that allows the respondent to record any other issues not covered by the questionnaire. This is one way of avoiding any frustration on the part of the respondent, as well as allowing them to express any thoughts, questions or concerns they might have. Lastly, there should be a message at the end thanking the respondents for their time and patience in completing the questionnaire.

Wording of questions

One of the most important factors in any survey is the design of the actual questionnaire. The questions and instructions should be easy to understand and respond to. The way a question is worded is very important as the same question worded in a different manner may achieve different results. Consider the following.

Abbreviations and acronyms

Always spell out the complete form of abbreviations and acronyms.

Example: Do you know if the pop figures are available online?

Better wording: Did you know that the population figures from the 2006 Census of Population are available on the Statistics Canada website at www.statcan.gc.ca?

Example: Have you ever participated in our LFS survey?

Better wording: Have you ever participated in a Labour Force Survey for Statistics Canada?

Complex words and terminology

Avoid specialized terminology and complicated words.

Example: Do you know who is leading the talks surrounding the impending amalgamation of surrounding constituencies into the "new metro" areas?

Better wording: Do you know who is leading the talks in each of the provinces regarding the amalgamation of cities, towns, villages and rural areas into "new metro" areas?

Example: Have you ever received a pneumococcus vaccination?

Better wording: Have you ever received a flu vaccination?

Frame of reference

Give all the details concerning the question's frame of reference.

Example: What is your income?

Does the word "your" refer to the respondent's personal income, family income or household income? Does the word "income" refer to salary and wages only, or does it include tips or income from other sources? Because there is no specific time period mentioned, does this question refer to last week's income, last month's or last year's income?

This question is too vague. It should be reworded so that all of the specific details concerning the frame of reference are given.

Better wording: What was your household's total income, from all sources before taxes and deductions, for last year?

Specific questions

A question's frame of reference is not the only specific detail required. In order to get a uniform response from the entire sample, the question sometimes needs to state the type of response needed.

Example: Respondents are shown a bottle of orange drink and are asked, “How much orange juice do you think this bottle contains?”

Some of the results from this question are outlined below:

- One orange and a little water and sugar
- 25% orange and 75% carbonated water
- Juice of one-half dozen oranges
- Three ounces of orange juice
- Full strength
- A quarter cup of orange juice
- None
- Not much
- Don’t know
- A pint
- Most of it
- About a glass and a half

Better wording: This bottle holds 250 millilitres (mL) of orange drink. How many millilitres of this drink would you say are orange juice?

Double-barreled questions

Examples:

Do you plan to leave your car at home and take the bus to work during the coming year?

Does your company provide training for new employees and retraining for existing staff?

Each of the above examples asks two questions rather than one:

In the first example, the question asks respondents if they plan to leave their cars at home, and whether or not they are taking the bus for the next year.

The second example asks respondents if their company provides training for new employees as well as providing retraining for existing employees.

In some instances, the answer to each half of the question is the same. However, sometimes there could be two very separate answers, which would make interpreting this question difficult.

The best solution could be to split such questions in two.

Loaded questions

The following examples demonstrate how a loaded question can impact the respondent’s results.

Example 1:

In your opinion, should Sunday shopping be allowed in Ontario; that is, should stores that want to stay open on Sunday be allowed to stay open on Sundays if they want to?

- Results:
    - ▶ 73% In favour of Sunday shopping
    - ▶ 25% Opposed to Sunday shopping
    - ▶ 2% No opinion

Example 2:

In your opinion, should a Sunday pause day be adopted in Ontario; that is, should the government make Sunday the one uniform day a week when most people do not have to work?

- Results:
    - ▶ 50% Opposed to a Sunday pause day
    - ▶ 44% In favour of a Sunday pause day
    - ▶ 6% No opinion

Source: Toronto Area Survey, 1991.

The wording of the first question asks whether the respondents were in favour of Sunday shopping, while the second question was worded to ask respondents whether they were in favour of not working on Sundays. As a result, there was a significant change in the data.

A possible explanation for the difference in the results could be that some respondents did not quite understand the implications of the question. Some people may be opposed to working on Sundays, but are still in favour of shopping. However, if no one works on Sundays, then stores cannot stay open for shoppers!

Open or closed questions

Generally, there are two types of questions: **open** and **closed**. Open questions give respondents an opportunity to answer the question in their own words. Closed questions give respondents a choice of answers and the respondent is supposed to select one.

Open question

What is the most important issue facing today's youth?

Closed question

Which of these is the most important problem facing today's youth?

- Unemployment
- National unity
- Environment
- Youth violence
- Rising tuition fees
- Drugs in schools
- Need for more computers in schools
- Career counseling

There are advantages and disadvantages to using one type of question versus another. The open question allows the respondent to interpret the question and answer it anyway he or she chooses. The respondent writes the answer or the interviewer records verbatim what the respondent says in answer to the question.

The closed question restricts the respondent to select an answer from the specified response options. For the respondent, a closed question is easier and faster to answer and for the researcher, closed questions are easier and less expensive to code and analyse. Also, closed questions provide consistency, an element that is not necessarily going to occur with an open question.

Questionnaire testing

This is a fundamental step in developing a questionnaire. Questionnaire testing allows:

- to discover poor wording or ordering of questions,
- to identify errors in the questionnaire layout and instructions,
- to determine problems caused by the respondent's inability or unwillingness to answer the questions,
- to suggest additional response categories that can be pre-coded on the questionnaire, and
- to provide a preliminary indication of the length of the interview and any refusal problems.

Testing can include the complete questionnaire or only a particular portion of it. The complete questionnaire will at some point in time have to be fully tested.

3.3.3 Role of interviewers

It is important to note that not all persons who collect data are interviewers. In some instances, people go into grocery stores or clothing stores to collect data. They manually record the prices from a list of goods and services on hand-held devices and then report their data to Statistics Canada staff.

However, the role of the interviewer is very important. The process of interviewing people to collect data involves a number of skills. Without these skills, the quality of data collected can be affected. Therefore, the skills sought in an interviewer are the following:

- good communication skills;
- a confident and professional appearance, and
- a driver's license, in the case of interviewers that will need to travel to do in-person interviews.

Statistics Canada employs a large number of interviewers to collect data. Interviewers are trained before collecting data. This training emphasizes that the interviewer's opening remarks and the manner in which they are made have a strong influence on a respondent's reaction and willingness to co-operate. Because of this, interviewers should ensure they carry out certain tasks before asking respondents to answer questions. They must:

- give the respondent their name and provide identification;
- explain that a survey is being conducted and by whom;
- describe the survey's purpose;
- explain that the respondent's household or business has been selected in the survey sample;
- give the respondent time to read or be informed about confidentiality issues, the voluntary or mandatory status of the survey, and any existing data-sharing agreements with other organizations, and
- read the introduction message of the questionnaire to the respondent.

In addition, it is important that the interviewer has appropriate skills and abilities such as:

- stimulating the respondent's interest;
- listening attentively;
- asking questions as worded for each respondent interviewed;
- NOT suggesting any answers to the respondent;

- answering the respondent's questions properly;
- keeping the respondent "on track", and
- explaining that the information collected is confidential.

Above all, the interviewer should let respondents know that he or she understands the respondent.

3.4 Processing

In the first half of this section, we have covered in detail the many steps involved in collecting data with a survey or census: the planning stage, the different ways to select a sample, how to reach selected units and to collect the information needed to meet the objective of the survey using a well-designed questionnaire and often with the help of interviewers. Once data has been obtained, either from this series of steps or from an administrative or alternative source of information, it's time to process it so it's ready to be used to produce statistical information.

This section describes the five processing steps that are commonly used to transform data prior to estimation. The complexity of the processing and the time and resources needed to prepare data depend on the type of source and the quality of data.

3.4.1 Coding

Coding is any process that assign a value (code) to a response. It means that coding entails either assigning a code to a given response or comparing the response to a set of codes and selecting the one that best describe the response. The code can be a numeric value or a character string. There could be different ways to do this translation, but alternative coding approaches affect the quality and cost of data produced.

Questionnaires usually have two types of questions—closed questions and open questions. The responses to these questions affect the type of coding performed. The following question is an example of a closed question:

To what degree is sport important in providing you with the following benefits?

<1/> Very important

<2/> Somewhat important

<3/> Not important

The following code structure is an example of an open question:

What sports do you participate in?

Specify ______________

In the case of closed questions, the response categories are determined before collection, with the numerical code usually appearing on the questionnaire beside each response category. For open questions, coding occurs after collection and may be either manual or automated. For some questions, coding may be straightforward (e.g. marital status). In other cases, such as geography, industry and occupation, a standard coding system is strongly recommended when available. But for many questions where no standard coding system exists, determining a good coding scheme is a non-trivial task.

Automated coding systems

Manual coding requires interpretation and judgment on the part of the coder, and may vary between coders. Due to advances in technology, resource constraints and, most importantly, concerns about timeliness and quality, coding is becoming more and more automated.

In general, two files are input to an automated coding system. One file contains either the survey responses or administrative files, which are to be coded, referred as the input file. The other file is called the reference file, which contains the predetermined code set. Then, for each record from the input file, a search is performed in the reference file. If a match is found, the code in the reference file is assigned to the corresponding record from the input file. Otherwise, the code is left blank. Some of the advantages of an automated coding system are that the process increasingly becomes faster, consistent, and more economical.

There are already many automated systems in use at Statistics Canada. For example, the Labour Force Survey data files are collected from the Regional Offices of Statistics Canada and are run through an automated coding system that assigns industry and occupation codes based on the North American Industrial Classification System (NAICS) and the National Occupation Classification (NOC). The rejected records (those that do not have a match with the written response) are the only data to be manually coded.

Recently, machine learning techniques have been used for Statistics Canada's Business Register to help assign industry codes by using business names and business addresses. This leads to improvement to the coverage of the Business Register, which is the sampling frame for the majority of business surveys in Statistics Canada, and, ultimately, to improvement to the data quality of many business surveys.

3.4.2 Capture

Data capture is any process that converts data into a format that can be used by a computer. In everyday life, there are a variety of devices used to capture data.

Let's say you wake up one morning and check your smartphone for the weather. You would use your phone's touch screen to select the appropriate app. Then, to check the news, you might open your browser and type in the name of your favorite news page using the on-screen keyboard. A little later, you go out to run some errands. You're wearing a smart watch on your wrist that uses sensors to track your step count and heart rate. At the checkout, the cashier will use a scanner to scan the barcode of each item that matches the name of the item and the corresponding price in the database. If you pay with a credit card, the cashier will hand you an electronic payment terminal. The terminal will take the information from your card using the chip it contains or a contactless payment system. You may be asked to enter your personal identification number (PIN) on the keypad. Back home, you decide to play a video game for entertainment. You use a remote control to select the game of your choice and a controller to control your actions in the game. Or you sit down at your laptop to answer your emails. You'll use the touchpad or a mouse to open the right software application and the keyboard to type your responses.

All of these devices allowed you to capture the data needed for your day's activities:

- Touch screen
- On-screen keyboard
- Watch with sensors
- Scanner
- Electronic payment terminal
- Remote control
- Gamepad
- Physical keyboard
- Touchpad
- Mouse

Camera and voice recorder of your smartphone could also be added to the list.

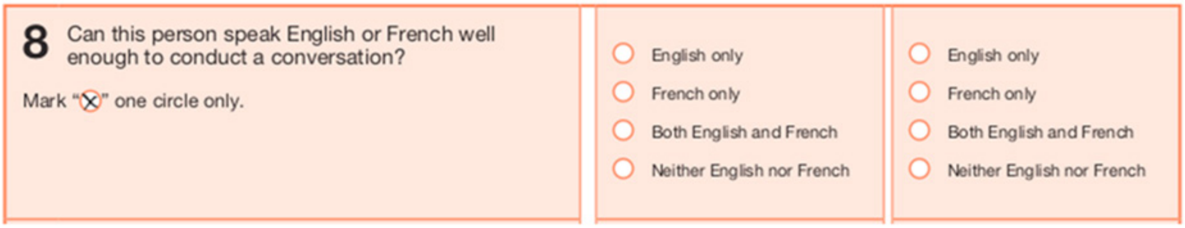
All of these devices and many others could potentially be used to capture data in a survey, depending on the information that needs to be collected. In the case of a survey that requires collecting data from respondents, an interviewer with a computer or the respondent him/herself, if provided with a self-completed electronic questionnaire, will enter the data using the mouse and keyboard. A smartphone or tablet computer could also be used for this purpose. Other means can be used to coordinate the collection, such as barcode or PIN systems that link the person selected from the survey frame to the data captured.

One aspect of data entry with which you may be less familiar is how to proceed when survey data are collected through a printed form. This may be more convenient and faster in some contexts, but requires more work to capture the data.

The capture can be done manually. This involves a person reading the answers on the printed form, coding them and entering them themselves using a computer keyboard. In this case, a good coding system is essential to speed up data capture and reduce the risk of errors.

Data entry can also be done automatically by scanning the questionnaires. In the case of closed questions, Optical Mark Recognition (OMR) can detect which boxes have been checked by a respondent. For example, figure 3.4.2.1 shows how for question 8 of the 2021 Census, “Can this person speak English or French well enough to conduct a conversation?”, the respondent must mark one circle corresponding to one of the four choices of answer.

Figure 3.4.2.1
Example of circles to be marked to answer to a closed question



8 Can this person speak English or French well enough to conduct a conversation?

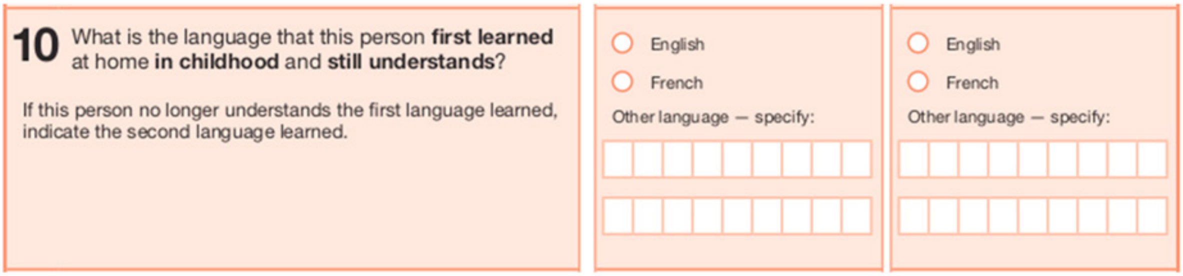
Mark “⊗” one circle only.

○ English only
○ French only
○ Both English and French
○ Neither English nor French

○ English only
○ French only
○ Both English and French
○ Neither English nor French

For open-ended questions, Intelligent Character Recognition (ICR) can help capture and identify the letters in each box. Figure 3.4.2.2 shows how for question 10 of the 2021 Census, “What is the language that this person first learned at home in childhood and still understands?’, the respondent must indicate the language by writing one letter in each of the cases provided for this purpose.

Figure 3.4.2.2
Example of boxes to be filled using molded letters to answer to an open question



10 What is the language that this person **first learned** at home **in childhood** and **still understands**?

If this person no longer understands the first language learned, indicate the second language learned.

○ English
○ French
Other language — specify:
□□□□□□□□□
□□□□□□□□□

○ English
○ French
Other language — specify:
□□□□□□□□□
□□□□□□□□□

Whether it is done manually or automatically, the input of responses to a printed form can be subject to errors. It is therefore essential to put processes in place to control quality.

3.4.3 Editing

In an ideal world, data would be collected without any errors. Unfortunately, responses, either from surveys or from administrative files, may be missing, incomplete or incorrect. Data editing is the application of checks to detect missing, invalid or inconsistent entries or to point to data records that are potentially in error. No matter what type of data you are working with, certain edits are performed at different stages or phases of data collection and processing. Data editing is described and illustrated here by focusing on surveys, but it is also widely applied to other data sources, such as administrative data, to ensure the data quality.

Data editing begins by asking the question, “What could be the causes for errors in our files?” There are several situations where errors can be introduced into the data, and the following list gives some of them:

- A respondent could have misunderstood a question.
- A respondent or an interviewer could have checked the wrong response.
- A coder could have miscoded or misunderstood a written response.
- An interviewer could have forgotten to ask a question or to record the answer.
- A respondent could have provided inaccurate responses.
- Some questions have been left blank.

Always keep in mind the objectives of data editing:

- to ensure the accuracy of data;
- to establish the consistency of data;
- to determine whether the data are complete;
- to ensure the coherence of aggregated data;
- to obtain the best possible data available.

Applying editing rules

So, how do we edit? The first step is to apply rules, or factors to be taken into consideration, to the data. These rules are determined by the expert knowledge of a subject-matter specialist, the structure of the questionnaire, the history of the data, and any other related surveys or data set.

Expert knowledge can come from a variety of sources. The specialist could be an analyst who has extensive experience with the type of data being edited. An expert could also be one of the survey sponsors who are familiar with the relationships between the data.

The layout and structure of the questionnaire will also impact the rules for editing data. For example, sometimes respondents are instructed to skip certain questions if the questions do not apply to them or their situation. This specification must be respected and incorporated into the editing rules.

Lastly, other data sources relating to the same sort of variables or characteristics are used in order to establish some of the rules for editing data. For example, business surveys usually collect financial data of businesses. The same information could be available from the tax returns of the company. Thus, the tax data can be used to develop editing rules for validating survey data.

Data editing types

There are several types of commonly used data editing, which include:

- **Validity edits** look at one question field or cell at a time. They check to ensure the record identifiers, invalid characters, and values have been accounted for; essential fields have been completed (e.g. no quantity field is left blank where a number is required); specified units of measure have been properly used; and the reported data lie within an allowed range of value (e.g. the reporting time is within the specified limits). In computer-assisted data collection, such as web surveys, real-time data editing is typically built into the data collection system so that the validity of the data is evaluated as the data are collected.
- **Duplication edits** examine one full record at a time. These types of edits check for duplicated records, making certain that a respondent or a survey unit has only been recorded once. A duplication edit also checks to ensure that the respondent does not appear in the survey universe more than once, especially if there has been a name change. Finally, it ensures that the data have been entered in the system only once.
- **Consistency edits** compare different answers from the same record to ensure that they are coherent with one another. For example, if a person is declared to be in the 0 to 14 age group, but also claims that he or she is retired, there is a consistency problem between the two answers. Inter-field edits are another form

of a consistency edit. These edits verify that if a figure is reported in one section, a corresponding figure is reported in another.

- **Historical edits** are used to compare survey answers in current and previous surveys. For example, any dramatic changes since the last survey will be flagged. The ratios and calculations are also compared, and any percentage variance that falls outside the established limits will be noted and questioned.
- **Statistical edits** look at the entire set of data. This type of edit is performed only after all other edits have been applied and the data have been corrected. The data are compiled and all extreme values, suspicious data and outliers are rejected.
- **Miscellaneous edits** fall in the range of special-reporting arrangements; dynamic edits particular to the survey; correct classification checks; changes to physical addresses, locations or contacts; and legibility edits (i.e. making sure the figures or symbols are recognizable and easy to read).

Data editing is influenced by the complexity of the questionnaire. Complexity refers to the length, as well as the number of questions asked. It also includes the detail of questions and the range of subject matter that the questionnaire may cover. In some cases, the terminology of a question can be very technical. For these types of surveys, special reporting arrangements and industry-specific edits may occur.

Data editing levels

Data editing can be performed manually, with the assistance of computer programming, or a combination of both techniques. Depending on the medium (electronic, paper) by which the data are submitted, there are two levels of data editing—**micro-** and **macro-editing**.

- Micro-editing corrects the data at the record level. This process detects errors in data through checks of the individual data records. The intent at this point is to determine the consistency of the data and correct the individual data records.
- Macro-editing also detects errors in data, but does this through the analysis of aggregate data (totals). The data are compared with data from other surveys, administrative files, or earlier versions of the same data. This process determines the comparability of data.

3.4.4 Imputation

Editing is of little value to the overall improvement of the actual survey results, if no corrective action is taken when items fail to follow the rules set out during the editing process. When all of the data have been edited using the applied rules and a file is found to have missing data, then imputation is usually done as a separate step.

Missing or invalid values definitely impact the quality of the survey results. Imputation is the process used to assign replacement values for missing, invalid or inconsistent values that have failed edits. This occurs after a follow-up with respondents (if possible), manual review and correction of questionnaires (if applicable). At this stage, all kinds of errors are corrected, including errors made by respondents and errors occurred during coding and data capture.

Imputation procedures are designed to fill in the gaps. In general, changes are made to the minimum number of fields until the completed record passes all of the edits. When these errors are detected, values for invalid, missing or incomplete entries are imputed or replaced with appropriate values, and answers are provided for non-response questions. This procedure is best accomplished by those with full access to the microdata and in possession of good auxiliary information.

Although imputation can improve the quality of the final data, care must be taken to choose an appropriate imputation methodology. Some methods of imputation do not preserve the relationship between variables. In fact, some can actually distort the underlying distributions.

Some commonly used methods of data imputation methods include:

- **Deductive imputation** is usually the first method used. This method is used when a value can be deducted with certainty and can be completed during the collection, capture, editing, or later stages of data processing. Deductive imputation is used when there is only one possible response to the question (e.g. all the values are given but the total or subtotal is missing).

- **Hot-deck imputation** uses the answers from another record of the same survey, referred to as the donor, to answer the question (or set of questions) that needs imputation. The donor can be randomly selected from a pool of donors with the same set of predetermined characteristics. For example, if a questionnaire has been returned with the yearly income missing, then we could determine donor characteristics as records with the same province, same occupation and same amount of experience as the respondent from the survey requiring imputation. A list of possible donors matching these criteria is created and one of them is randomly selected. Once a donor is found, the donor response (in this case, the yearly income) replaces the missing or invalid response.
- **Cold-deck imputation** is similar to hot-deck imputation. The difference is that hot-deck imputation uses donors from the same survey, while code-deck imputation uses donors from another source, such as historical data from an earlier iteration of the same survey or from administrative data.
- **Mean value imputation** is to replace the missing or inconsistent value by the mean value calculated from the responding units with the same set of predetermined characteristics. For example, if a record is missing a total number for an individual's yearly income, then we could impute the observed average income in that individual's province for the same occupation with the same level of experience as the respondent. One drawback of mean imputation is that it destroys distribution and the relationships between variables by creating an artificial spike at the group mean. This artificially lowers the estimated sampling variance of the final estimates if conventional formulas for the sampling variance are used.
- **Nearest neighbor imputation** is another type of donor imputation. In this case, some sort of criteria must be developed to determine the responding unit the most similar to the unit with the missing value in accordance with the predetermined characteristics. The closest unit to the missing value is then used as the donor.

There are other more sophisticated imputation methods, which use statistical modelling to assign an appropriate replacement value.

The method of imputation can vary from survey to survey and even, depending on particular circumstances, within the same survey. Quite often, different methods are combined together to provide the most suitable value for a variable. These methods can be applied either manually or with the use of an automated system. To help facilitate this, Statistics Canada has developed a generalized imputation system to impute data based on the methodological input of experienced statisticians who have analyzed the survey and suggested the best approaches to impute meaningful data.

Although imputation can improve the quality of the final data, care should be taken when choosing an appropriate imputation methodology. One risk with imputation is that it can destroy reported data to create records that fit preconceived models that may later turn out to be incorrect. The suitability of the imputation methods depends upon the survey, its objectives, available auxiliary information and the nature of the error.

In addition, all the imputation methods can be applied to other sources of data, not just limited to survey data. For example, Statistics Canada receives and uses financial data from Canada Revenue Agency to reduce response burden, and these administrative data often have missing or inconsistent values. In order to make good use of them, some rigorous editing and imputation systems have been put in place to improve the data quality before moving to the next step.

Note also that in the case of total nonresponse, when very little or no data have been collected for a record or unit, a common approach is to perform a nonresponse weight adjustment, which will be discussed in detail later in the section on estimation.

3.4.5 Record linkage

Record Linkage is the process in which records or units from different data sources are joined together into a single file using non-unique identifiers, such as names, date of birth, addresses and other characteristics. It is also known as data matching, data linkage, entity resolution and many other terms depending on the fields it's been used. The initial idea of record linkage goes back to the 1950s, then this technique has been applied by people from a wide range of areas, such as data warehousing and business intelligence, historical research, and medical practice and research.

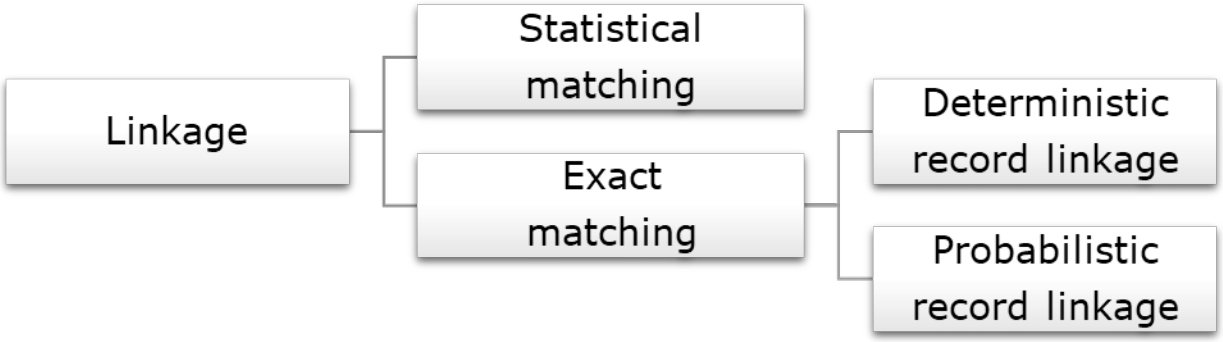
Matching has a long history of uses in statistical surveys and administrative data development. In Statistics Canada, record linkage is used in creating a sampling frame, removing duplicates from files, providing extra information to assist data processing, or combining files so that relationships on two or more data elements from separate files can be studied. For example:

- A business register consisting of names, addresses, and other identifying information such as complete financial information might be constructed from tax and employment databases.
- A survey of retail establishments or agricultural establishments might combine results from an area frame and a list frame. To produce a combined estimator, units from the area frame would need to be identified in the list frame.
- The coverage of the Census of the population can be measured by linking the Census records to other sources of administrative data and by estimating the percentage of individuals found in one source but not in the other source.

Types of Linkage

There are two types of record linkage: exact matching and statistical matching. Exact matching can be divided into two subtypes: deterministic record linkage and probabilistic record linkage, as illustrated by figure 3.4.5.1 below.

Figure 3.4.5.1
Types of record linkage



Statistical Matching

The purpose of statistical matching is to create a file reflecting the underlying population distribution. Records that are combined do not necessarily correspond to the same entity, such as a person or a business. The files that are being matched can have different units but referring to the same population. It is assumed that the relationship of the variables in the population will be similar to the relationship on the files. This method is mainly used in market research and seldom used by official statistical agencies.

Exact Matching

The goal of exact matching is to link information about a particular record in one file to information on a secondary file in order to create a single file with correct information for each record. The linkage is performed at the level of record, such as a link of mortality records to the population census.

Deterministic Record Linkage

This is the simplest form of record linkage, which produces links based on common identifiers or variables among the available data sources. It is often the case that no single variable exists that is free of errors, presents on the majority of data and has enough discrimination power. Only a combination of variables will be able to discriminate between two records. This is one technique often used by official statistical agencies. Statistics Canada uses this method for building its business, address and population registers, which involve in multiple survey operations subsequently.

Probabilistic Record Linkage

This is another type of exact matching. Like in the other case, there is no unique identifier available for matching. Unlike the deterministic matching, probabilistic can compensate if the information is incomplete or/and subject to error. Records, which are not in complete agreement for each variable, can be linked together to build a set of potential pairs. A score is then calculated for each potential pair. After that, a linkage status is assigned to each potential pair based on the score.

Remark

There are numerous factors to consider to determine which type of record linkage to use, such as the purpose of the linkage, type of data, cost, time, confidentiality, acceptable precision level and type of error. In general, deterministic matching is less computer intensive but it involves more manual intervention. Probabilistic linkage is more time consuming and computer intensive, and will require specialized software. However, it generally produces more reliable results than deterministic linkage.

3.5 Estimation

As we now know, the goal of collecting data, including conducting surveys or acquiring data from other sources, is to obtain information about a particular population. When the sample has been selected and the information collected (see the data collection section) and data processed (see the data processing section), there still remains the task of linking the information gathered from the sample back to the target population.

Estimation is the process of finding an estimate (or approximation), which is a value derived from the best data available. Typically, estimation involves using the value derived from a sample to estimate the value of a corresponding population characteristic. Researchers are usually interested in looking at estimates of many statistics—totals, averages and proportions being the most frequent—for different variables. For example, a sample survey could be used to produce any of the following statistics:

- the proportion of smokers among all people aged 15 to 24 in the population;
- the average earnings of men and women with a university degree;
- the total number of cars possessed by the whole survey population.

In this section, we will outline what is the estimation process, starting with weighting followed by the estimation of sampling error and a description of the different sources of non-sampling error. Like sampling, estimation requires advanced knowledge of mathematical statistics. Examples presented in this section are based on the simplest sampling design, simple random sampling, and only aims to give an overview of the estimation process.

3.5.1 Weighting

The principle behind estimation in a probability survey is that each sample unit represents not only itself, but also several units of the survey population. The design weight of a unit usually refers to the average number of units in the population that each sampled unit represents. This weight is determined by the sampling method and is an important part of the estimation process.

While the design weights can be used for estimation, most surveys produce a set of estimation weights by adjusting the design weights to improve the precision of the final estimates. The two most common reasons for making adjustments are to account for nonresponse and to make use of pertinent data available from other sources. Once the final estimation weights have been calculated, they are applied to the sample data in order to compute estimates.

Design weight

The first step in estimation is assigning a weight to each sampled unit. The **design weight** (w_d), which is the average number of units in the population that each sampled unit represents, is the inverse of its inclusion probability (π) in the sample.

$$w_d = 1/\pi$$

If the inclusion probability is 1/50, then each selected unit represents on average 50 units in the population and the design weight is $w_d = \mathbf{50}$.

Some sample designs assign the same design weights for all units in the sample, while others give different design weights to sampled units for various reasons, such as improving precision or reducing cost.

Example 1: Simple Random Sample

Suppose there are *N* =100 Grade 12 (or secondary 5) students in a high school. A simple random sample of size *n* =25 students is selected, and the selected students are invited to complete a questionnaire about their career plan.

- The inclusion probability is:
  $\pi = n / N = 25 / 100 = 1 / 4$.
- The design weight is:
  $$w_d = {}^{1}\!/_{\pi} = 1 / \frac{1}{4} = 4$$.

Each student selected in the simple represents four students of the school.

Production of simple estimates

Estimates can be produced after weights are calculated while only simple estimates, such as totals, averages and proportions, are covered here.

Estimating a population total

The estimate of the total number ($\hat{Y}$) of units in the population is calculated by multiplying the weight and the value of interest for each selected unit then summed over all in sample units. For categorical variables, the estimate is actually calculated by adding together the weights of the responding units.

Example 2: Simple Random Sample (Continued)

Suppose that within the 25 students selected in the sample, there are about 10 applied to science programs. Then, the total number of students applied to science programs is:

$$\hat{Y} = 4 \times 10 = 40$$

Estimating a population average

The estimate of the average ($\hat{\bar{Y}}$) in the population is the estimate of the total value for the variable in interest ($\hat{Y}$) divided by the estimate of the total number of units ($\hat{N}$) in the population.

$$\hat{\bar{Y}} = \frac{\hat{Y}}{\hat{N}}$$

Example 3: Simple Random Sample (Continued)

Usually, students apply to more than one program when applying for university study. Suppose that within the 25 students selected in the sample, 5 of them apply to only 1 program, 10 of them apply to 2 programs and 10 of them apply to 3 programs. Then, the average number of applications per student is calculated as following:

- Total number of applications is given by:

$$\hat{Y} = \left(4 \times 5 \times 1\right) + \left(4 \times 10 \times 2\right) + \left(4 \times 10 \times 3\right) = 220$$

- Total number of students is given by:

$$\hat{N} = 4 \times 25 = 100$$

- Average number of applications per student is given by:

$$\widehat{\overline{Y}} = \frac{\hat{Y}}{\hat{N}} = \frac{220}{100} = 2.2$$

Estimating a population proportion

The estimate of the proportion in the survey population having a given characteristic is quite similar as estimating a population average in terms of the mathematical formula. It is also calculated as a quotient between two estimated totals. The main difference is the numerator, which indicates the estimate of the total number of units possessing the given characteristic (c) when estimating a proportion ($\hat{p}$). However, the numerator is the estimate of the total value for quantitative data when estimating an average.

$$\hat{P} = \frac{\widehat{Nc}}{\hat{N}}$$

Example 4: Simple Random Sample (Continued)

Suppose within the 25 students selected in the sample, there are 10 females and 15 males. Overall, 10 students apply for science programs with 5 females and 5 males. The proportion of students apply for science programs by gender is calculated as following:

- Total number of students applied science programs by gender is given by:

$$\hat{N}_{male,science} = 5 \times 4 = 20$$

$$\hat{N}_{female,science} = 5 \times 4 = 20$$

- Total number of students by gender is given by:

$$\hat{N}_{male} = 15 \times 4 = 60$$

$$\hat{N}_{female} = 10 \times 4 = 40$$

- Proportion of students applied science programs by gender is given by:

$$\hat{P}_{male,science} = \frac{\hat{N}_{male,science}}{\hat{N}_{male}} = \frac{20}{60} = 1/3$$

$$\hat{P}_{female,science} = \frac{\hat{N}_{female,science}}{\hat{N}_{female}} = \frac{20}{40} = 1/2$$

Other estimation methods

The estimation method described above for Simple Random Sampling is the simplest estimation method, and there are other more advanced ones available, which are widely applied in many surveys. The most appropriate estimation method to use is determined by a few factors, such as the characteristics to be estimated, the different types of data, reliability, cost and timeliness, etc. At Statistics Canada, specialized estimation systems are used to produce estimates involving complicated procedures in a timely manner.

Adjusting the weights

Quite often design weights have to be adjusted prior to estimation, and there are two main types of adjustment: nonresponse adjustment and adjustment for external information.

Adjusting for nonresponse

Almost all surveys suffer from nonresponse, which occurs when all or some key information requested from sampled units is unavailable for some reason, such as the sample unit refuses to participate, no contact is made, the unit cannot be located or the information obtained is unusable. The easiest way to deal with such nonresponse is to ignore it, but this leads to inaccurate estimates.

Two common ways of dealing with this kind of nonresponse is to impute missing answers or to adjust the design weights based on the assumption that the responding units represent both responding and nonresponding units. The design weights of the non-respondents are then redistributed among the respondents.

Adjusting for external information

Sometimes information about the survey population is available from other sources, for example information from a census or an administration file. This information can also be incorporated in the weighting process.

There are two main reasons for using external (auxiliary) data at estimation. The first reason is that it is often important for the survey estimates to match known population totals or estimates from another, more reliable, survey. For example, many social surveys adjust their survey estimates in order to be consistent with estimates (age, sex distributions, etc.) of the most recent census of the population. External information may also be obtained from administrative data or from another survey that is considered to be more reliable because of its larger sample size or because the published estimates must be respected.

The second reason is to improve the precision of the estimates, as long as the values of the auxiliary variables are collected for the surveyed units and that population totals or estimates are available for these variables from another reliable source.

3.5.2 Sampling error

An important part of estimation is estimating the magnitude of the sampling error in the estimate. This provides a measure of the precision of the survey's estimates for the specific sample design. Sampling error can only be estimated if probability sampling is used.

The sampling error is the error caused by observing a sample instead of the whole population. It arises from estimating a population characteristic by looking at only one portion of the population rather than the entire population, and refers to the difference between the estimate derived from a sample survey and the true value that would result if a census of the whole population were taken under the same conditions. There is no sampling error in a census because the calculations are based on the entire population.

Estimating the sampling error

As mentioned before, any estimates derived from samples are subject to sampling error because only a part of the population was observed. A different sample could have come up with different estimates. Sampling error causes variability among estimates derived from different samples when keeping the same sample size and design, and the same estimation method used. It's measured commonly by the sampling variance, which depends on many things, including the sampling method, the estimation method, the sample size and the variability of the estimated characteristic.

Sampling variance

In simple sample designs, such as Simple Random Sampling, the sampling variance can be calculated directly using a formula. However, formula usually doesn't exist for more complex designs. In this case, an estimate of the sampling variance can be calculated using methods such as Taylor linearization or resampling methods such as the jackknife and the bootstrap.

Regardless of which method is used for variance estimation, it has to incorporate sample design properties such as stratification, clustering and multistage or multi-phase selection, if applicable.

Other factors affecting the magnitude of the sampling variance include the following:

- In general, sampling variance decreases as **sample size** increases but the change is not proportional.
- **Population size** has an impact on the sample variance for small to moderate sized population. For large populations, its impact is minor.
- **Variability of the characteristic of interest in the population** also affects the size of the sampling error. The greater the difference between the population units, the larger the sample size required to achieve a specific level of precision.
- A **sampling plan**, which includes a sample design and an estimation procedure, also affect the magnitude of sampling error. The method of sampling, called "sample design," can greatly affect the size of the sampling error. Surveys involving a complex sample design could lead to larger sampling error than a simpler one. The estimation procedure also has a major impact on the sampling error. These concepts are examined in greater detail in the section on sampling.

Other measures of sampling error

Except using sampling variance to measure sampling error, other frequently used methods also exist, including standard error, coefficient of variation, margin of error and confidence interval.

The **standard error** is the square root of the sampling variance. This measure is easier to interpret since it provides an indication of sampling error using the same scale as the estimate whereas the variance is based on squared differences.

The **coefficient of variation (CV)** assesses the size of the standard error relative to the estimate of the characteristic being measured. It is the ratio of the standard error of the estimate to the average value of the estimate itself. CV is very useful in comparing the precision of sample estimates, where their sizes or scale differ from one another. Even though CV is widely used in Statistics Canada's official releases, it's not recommended for measuring the precision for proportions, especially when the estimated proportions are close to 0 or 1. In this case, confidence interval is more appropriate to use.

The **confidence interval (CI)** gives a range of values around the estimate that is likely to include the unknown population value with a given probability. This probability is the confidence level of the IC. For a given estimate in a given sample, using a higher confidence level generates a wider CI, meaning a less precise CI. The most commonly used confidence level is 95%, but confidence levels of 99% or 90% are also used in certain circumstances.

The **margin of error** is half the width of the CI. The larger the margin of error, the less confidence one should have that a result would reflect the result of a survey of the entire population. It is often used to report sampling error by pollsters or journalists.

Example 1

It is common to see the results of a survey reported in a newspaper as follows:

> According to a recent survey, **15%** of Ottawa residents attend religious services every week. The results, based on a sample of 1,345 residents, are considered accurate within **plus or minus 3 percentage points 19 times out of 20**.

In this example, the expression "19 times out of 20" means that if the survey was repeated many times, then the confidence interval would cover the true population value 19 times out of 20. This is equivalent to a 95% confidence level. The expression "plus or minus 3 percentage points" means that the margin of error is 3%. Therefore, the value of the estimation is 15% and the corresponding 95% CI is 12% to 18%.

3.5.3 Non-sampling error

Non-sampling error refers to all sources of error that are unrelated to sampling. Non-sampling errors are present in all types of survey, including censuses and administrative data. They arise for a number of reasons: the frame may be incomplete, some respondents may not accurately report data, data may be missing for some respondents, etc.

Non-sampling errors can be classified into two groups: **random errors** and **systematic errors**.

- **Random errors** are errors whose effects approximately cancel out if a large enough sample is used, leading to increased variability.
- **Systematic errors** are errors that tend to go in the same direction, and thus accumulate over the entire sample leading to a bias in the final results. Unlike random errors, this bias is not reduced by increasing the sample size. Systematic errors are the principal cause of concern in terms of a survey's data quality. Unfortunately, non-sampling errors are often extremely difficult, if not impossible, to measure.

Types of non-sampling error

Non-sampling error can occur in all aspects of the survey process, and can be classified into the following categories: coverage error, measurement error, nonresponse error and processing error.

Coverage error

Coverage error consists of omissions (undercoverage), erroneous inclusions, duplications and misclassifications (overcoverage) of units in the survey frame. Since it affects every estimate produced by the survey, they are one of the most important types of error. In the case of a census, it may be the main source of error. Coverage error can have both spatial and temporal dimensions, and may cause bias in the estimates. The effect can vary for different subgroups of the population. This error tends to be systematic and is usually due to under coverage, which is why it's important to reduce it as much as possible.

Measurement error

Measurement error, also called **response error**, is the difference between measured values and true values. It consists of bias and variance, and it results when data are incorrectly requested, provided, received or recorded. These errors may occur because of inefficiencies with the questionnaire, the interviewer, the respondent or the survey process.

- **Poor questionnaire design**
  It is essential that sample survey or census questions are worded carefully in order to avoid introducing bias. If questions are misleading or confusing, then the responses may end up being distorted.
- **Interviewer bias**
  An interviewer can influence how a respondent answers the survey questions. This may occur when the interviewer is too friendly or aloof or prompts the respondent. To prevent this, interviewers must be trained to remain neutral throughout the interview. They must also pay close attention to the way they ask each question. If an interviewer changes the way a question is worded, it may impact the respondent's answer.
- **Respondent error**
  Respondents can also provide incorrect answers. Faulty recollections, tendencies to exaggerate or underplay events, and inclinations to give answers that appear more socially acceptable are several reasons why a respondent may provide a false answer.
- **Problems with the survey process**
  Errors can also occur because of a problem with the actual survey process. Using proxy responses, meaning taking answers from someone other than the respondent, or lacking control over the survey procedures are just a few ways of increasing the risk of response errors.

Non-response error

Estimates obtained after nonresponse has been observed and imputation has been used to deal with this nonresponse are usually not equivalent to the estimates that would have been obtained had all the desired values been observed without error. The difference between these two types of estimates is called the **nonresponse error**. There are two types of non-response errors: total and partial.

- **Total nonresponse error** occurs when all or almost all data for a sampling unit are missing. This can happen if the respondent is unavailable or temporarily absent, the respondent is unable to participate or refuses to participate in the survey, or if the dwelling is vacant. If a significant number of sampled units do not respond to a survey, then the results may be biased since the characteristics of the non-respondents may differ from those who have participated.
- **Partial nonresponse error** occurs when respondents provide incomplete information. For certain people, some questions may be difficult to understand, they may refuse or forget to answer a question. Poorly designed questionnaire or poor interviewing techniques can also be reasons which result partial nonresponse error. To reduce this form of error, care should be taken in designing and testing questionnaires. Adequate interviewer training and appropriate edit and imputation strategies will also help minimize this error.

Processing error

Processing error occurs during data processing. It includes all data processing activities after collection and prior to estimation, such as errors in data capture, coding, editing and tabulation of the data as well as in the assignment of survey weights.

- **Coding errors** occur when different coders code the same answer differently, which can be caused by poor training, incomplete instructions, variance in coder performance (i.e. tiredness, illness), data entry errors, or machine malfunction (some processing errors are caused by errors in the computer programs).
- **Data capture errors** result when data are not entered into the computer exactly as they appear on the questionnaire. This can be caused by the complexity of alphanumeric data and by the lack of clarity in the answer provided. The physical layout of the questionnaire itself or the coding documents can cause data capture errors. The method of data capture, manual or automated (for example, using an optical scanner), can also result in errors.
- **Editing and imputation errors** can be caused by the poor quality of the original data or by its complex structure. When the editing and imputation processes are automated, errors can also be the result of faulty programs that were insufficiently tested. The choice of an inappropriate imputation method can introduce bias. Errors can also result from incorrectly changing data that were found to be in error, or by erroneously changing correct data.

3.6 Quality management

Quality is an essential element at all levels of processing. Statistics Canada's reputation as the best statistical agency in the world is based on the quality of its data. To ensure the quality of a product or service in our survey development activities, both **quality assurance** and **quality control** methods are employed.

Quality assurance

Quality assurance refers to all planned activities necessary in providing confidence that a product or service will satisfy its purpose and the users' needs. In the context of survey conducting activities, this can take place at any of the major stages of survey development: planning, design, implementation, processing, evaluation and dissemination.

Examples of planned activities include

- improving a survey frame,
- changing the sample design,
- modifying the data collection process,
- improving follow-up routines,
- changing the processing procedures,
- revising the design of the questionnaire.

Quality assurance attempts to move quality upstream by anticipating problems before they occur and aims at ensuring quality via the use of prevention and control techniques.

Quality control

Quality control is a regulatory procedure through which we

- measure quality;
- compare quality with pre-set standards, and
- react to the differences between measured values and established standards.

Some examples of this include controlling the quality of the coding operation, the quality of the survey interviewing, and the quality of the data capture.

The objective of quality control is to achieve a given quality level with minimum cost. Some assurance and control functions are often performed within the survey unit itself, especially in connection with the tasks of data coding, capture and editing. Several of these procedures are automated, some partially automated and others employ purely manual methods.

Differences between quality assurance and quality control

Quality assurance

- anticipates problems before they occur
- uses all available information to generate improvements
- is not tied to a specific quality standard
- is applicable mostly at the planning stage
- is all-encompassing in its activities.

Quality control

- responds to observed problems
- use ongoing measurements to make decisions on the processes or products
- requires a pre-specified quality standard for comparability
- is applicable mostly at the processing stage
- is a set procedure that is a subset of quality assurance.

3.7 Exercises

1. Poplar Ridge Academy has been given a sizeable grant: enough to build either a new library or gymnasium. But, as there is only money enough to build one facility, the principal wants to ask students which one they feel is in greater need of renovation.

The table below indicates the number of students by gender, per grade, from preschool to Grade 12 (secondary 5).

Table 3.7.1
Number of students by sex and grade level, Poplar Ridge Academy

	Preschool	1	2	3	4	5	6	7	8	9	10	11	12
Boys	9	8	9	9	13	20	23	28	78	74	69	71	60
Girls	6	8	11	10	13	18	35	34	63	62	61	88	70
Total	**15**	**16**	**20**	**19**	**26**	**38**	**58**	**62**	**141**	**136**	**130**	**159**	**130**

a. What is the total population of Poplar Ridge Academy?

b. The principal wants to sample 50% of the students. How many students would this be?

c. The principal wants to keep the correct proportion of girls to boys in the sample. Using the following formula, calculate the number of male preschool students that should be included in the sample.

$$\frac{number\ of\ male\ preschool\ students}{number\ of\ total\ students} \times size\ of\ sample\ survey$$

d. What type of sampling technique has been used here?

e. If the principal wishes to sample 180 students, how many boys and girls per grade should be surveyed? Put your answers in a table. (Results should be rounded to the nearest whole number.)

2. From what you already know about the statistical process, place the following steps in the correct order:

- processing
- collection
- information
- data

3. If data editing did not take place, what effect might this have on information produced?

3.8 Answers

1. a. The total student population of the Poplar Ridge Academy is 950.
   b. A sample of 50% of the school's student population would equal 475 students.
   c. The number of male preschool students to be included in a sample of 475 is 4 or 5.
   9/950 x 475 = 4.5
   d. The sampling method used here is stratified sampling.
   e. The following table features the breakdown of the 180 member sample so that it is proportionate to gender by grade level.

Table 3.8.1
Number of students by sex and grade level needed to make a sample of 180 students, Poplar Ridge Academy

	Preschool	1	2	3	4	5	6	7	8	9	10	11	12
Boys	2	2	2	2	2	4	4	5	15	14	13	13	11
Girls	1	2	2	2	2	3	7	6	12	12	12	17	13
Total	**3**	**4**	**4**	**4**	**4**	**7**	**11**	**11**	**27**	**26**	**25**	**30**	**24**

2. The correct order for these items is

- data
- collection
- processing
- information

3. Unedited data may contain errors or miscalculations, thereby causing the information to be wrong or incomplete. This inaccurate information requires editing before being released to the public.

4 Data exploration

At many steps of the process of producing statistical information, it can be useful to explore the data. It can be when evaluating if a source of data meets your needs, when you receive the raw data and want to decide what data processing steps will be needed to be able to use it or before proceeding with more advanced statistical analyses. No matter what the source of the data is, it is important to understand the data and identify limitations. For this purpose, you can ask yourself the following questions:

- What metadata is available for this data set? Are the descriptions of variables provided?
- What are the observed population, the observation unit and the reference period?
- Is it microdata or aggregate data?
- What are the types of the variables in the file?
- What are the frequency distributions of these variables? What are the measures of central tendency and dispersion?

This section starts with the presentation of some software applications that are useful for data exploration. Then the different types of variables are presented, followed by the descriptive statistics used to explore data, such as frequency tables and measures of central tendency and dispersion.

4.1 Data exploration tools

Software applications for charts, programming, databases and spreadsheets are commonly used to explore data. Here are some examples of applications:

- **Spreadsheets** are programs that allow adding columns and rows of figures, to calculate means and to perform descriptive statistical analyses. They can be used to create summaries of results as well as charts and graphs to better understand relations between variables. These can be displayed in a number of ways: bar charts, line charts, and pie charts are just a few examples of the data visualizations that can be produced.
- **Data is sometimes stored in databases** for easy access and to allow the production of summaries, aggregate data or reports. A database program should be able to store, retrieve, sort and analyze data.
- **Specialized programs** can be developed to edit, clean, impute and process the final tabular output. They offer the full service in one module and can be used each time the same survey is completed and entered within the system. These programs will produce results ready to be published.
- **Statistical software applications** are used for data processing and to produce summaries and data visualizations, but they can also be used to carry advanced statistical analyses such as modelling.

One example of a very popular data exploration tool is R software. R is a programming language and open-source software that anyone can download and install on their personal computer to transform, explore and analyze data. All charts and graphs presented in the upcoming sections were created with R.

Computer output obtained from these data exploration tools may be used in a variety of ways. They can be saved for future retrieval and use, be sent to other teams in electronic files or be disseminated online to communicate statistical information to users. Output is usually governed by the need to communicate specific information to a specific audience. To help determine the best output type for the information you have produced, ask yourself these questions:

- For whom is the output being produced?
- How the audience will best understand it?

4.2 Types of variables

A variable is a characteristic that can be measured and that can assume different values. Height, age, income, province or country of birth, grades obtained at school and type of housing are all examples of variables. Variables may be classified into two main categories: categorical and numeric. Each category is then classified in two subcategories: nominal or ordinal for categorical variables, discrete or continuous for numeric variables. These types are briefly outlined in this section.

Categorical variables

A categorical variable (also called qualitative variable) refers to a characteristic that can't be quantifiable. Categorical variables can be either nominal or ordinal.

Nominal variables

A nominal variable is one that describes a name, label or category without natural order. Sex and type of dwelling are examples of nominal variables. In Table 4.2.1, the variable "mode of transportation for travel to work" is also nominal.

Table 4.2.1
Method of travel to work for Canadians

Mode of transportation for travel to work	Number of people
Car, truck, van as driver	9,929,470
Car, truck, van as passenger	923,975
Public transit	1,406,585
Walked	881,085
Bicycle	162,910
Other methods	146,835

Ordinal variables

An ordinal variable is a variable whose values are defined by an order relation between the different categories. In Table 4.2.2, the variable "behaviour" is ordinal because the category "Excellent" is better than the category "Very good," which is better than the category "Good," etc. There is some natural ordering, but it is limited since we do not know by how much "Excellent" behaviour is better than "Very good" behaviour.

Table 4.2.2
Student behaviour ranking

Behaviour	Number of students
Excellent	5
Very good	12
Good	10
Bad	2
Very bad	1

It is important to note that even if categorical variables are not quantifiable, they can appear as numbers in a data set. Correspondence between these numbers and the categories is established during data coding. To be able to identify the type of variable, it is important to have access to the metadata (the data about the data) that should include the code set used for each categorical variable. For instance, categories used in Table 4.2.2 could appear as a number from 1 to 5: 1 for "very bad," 2 for "bad," 3 for "good," 4 for "very good" and 5 for "excellent."

Numeric variables

A numeric variable (also called quantitative variable) is a quantifiable characteristic whose values are numbers (except numbers which are codes standing up for categories). Numeric variables may be either continuous or discrete.

Continuous variables

A variable is said to be continuous if it can assume an infinite number of real values within a given interval. For instance, consider the height of a student. The height can't take any values. It can't be negative and it can't be higher than three metres. But between 0 and 3, the number of possible values is theoretically infinite. A student may be 1.6321748755 … metres tall. In practice, the methods used and the accuracy of the measurement instrument will restrict the precision of the variable. The reported height would be rounded to the nearest centimetre, so it would be 1.63 metres. The age is another example of a continuous variable that is typically rounded down.

Discrete variables

As opposed to a continuous variable, a discrete variable can assume only a finite number of real values within a given interval. An example of a discrete variable would be the score given by a judge to a gymnast in competition: the range is 0 to 10 and the score is always given to one decimal (e.g. a score of 8.5). You can enumerate all possible values (0, 0.1, 0.2…) and see that the number of possible values is finite: it is 101! Another example of a discrete variable is the number of people in a household for a household of size 20 or less. The number of possible values is 20, because it's not possible for a household to include a number of people that would be a fraction of an integer like 2.27 for instance.

4.3 Frequency distribution

The **frequency (f)** of a particular value is the number of times the value occurs in the data. The **distribution** of a variable is the pattern of frequencies, meaning the set of all possible values and the frequencies associated with these values. Frequency distributions are portrayed as frequency tables or charts.

Frequency distributions can show either the actual number of observations falling in each range or the percentage of observations. In the latter instance, the distribution is called a **relative frequency distribution**.

Frequency distribution tables can be used for both categorical and numeric variables. Continuous variables should only be used with class intervals, which will be explained shortly.

Let's look at some examples of frequency distribution and relative frequency distribution for discrete variables.

Example 1 – Constructing a frequency distribution table

A survey was taken on Maple Avenue. In each of 20 homes, people were asked how many cars were registered to their households. The results were recorded as follows:

1, 2, 1, 0, 3, 4, 0, 1, 1, 1, 2, 2, 3, 2, 3, 2, 1, 4, 0, 0

Use the following steps to present this data in a frequency distribution table.

1. Divide the results (***x***) into intervals, and then count the number of results in each interval. In this case, the intervals would be the number of households with no car (0), one car (1), two cars (2) and so forth.
2. Make a table with separate columns for the interval numbers (the number of cars per household), the tallied results, and the frequency of results in each interval. Label these columns **Number of cars, Tally** and **Frequency**.
3. Read the list of data from left to right and place a tally mark in the appropriate row. For example, the first result is a 1, so place a tally mark in the row beside where 1 appears in the interval column (**Number of cars**). The next result is a 2, so place a tally mark in the row beside the 2, and so on. When you reach your fifth tally mark, draw a tally line through the preceding four marks to make your final frequency calculations easier to read.
4. Add up the number of tally marks in each row and record them in the final column entitled **Frequency.**

Your frequency distribution table for this exercise should look like this:

Table 4.3.1
Frequency table for the number of cars registered in each household

Number of cars (x)	Frequency (f)
0	4
1	6
2	5
3	3
4	2

0 true zero or a value rounded to zero

By looking at this frequency distribution table quickly, we can see that out of 20 households surveyed, 4 households had no cars, 6 households had 1 car, etc.

Example 2 – Constructing a cumulative frequency distribution table

A **cumulative frequency distribution** table is a more detailed table. It looks almost the same as a frequency distribution table but it has added columns that give the cumulative frequency and the cumulative percentage of the results, as well.

At a recent chess tournament, all 10 of the participants had to fill out a form that gave their names, address and age. The ages of the participants were recorded as follows:

36, 48, 54, 92, 57, 63, 66, 76, 66, 80

Use the following steps to present these data in a cumulative frequency distribution table.

1. Divide the results into intervals, and then count the number of results in each interval. In this case, intervals of 10 are appropriate. Since 36 is the lowest age and 92 is the highest age, start the intervals at 35 to 44 and end the intervals with 85 to 94.
2. Create a table similar to the frequency distribution table but with three extra columns.
    - In the first column or the **Lower value** column, list the lower value of the result intervals. For example, in the first row, you would put the number 35.
    - The next column is the **Upper value** column. Place the upper value of the result intervals. For example, you would put the number 44 in the first row.
    - The third column is the **Frequency** column. Record the number of times a result appears between the lower and upper values. In the first row, place the number 1.
    - The fourth column is the **Cumulative frequency** column. Here we add the cumulative frequency of the previous row to the frequency of the current row. Since this is the first row, the cumulative frequency is the same as the frequency. However, in the second row, the frequency for the 35–44 interval (i.e., 1) is added to the frequency for the 45–54 interval (i.e. 2). Thus, the cumulative frequency is 3, meaning we have 3 participants in the 34 to 54 age group.
      1 + 2 = 3
    - The next column is the **Percentage** column. In this column, list the percentage of the frequency. To do this, divide the frequency by the total number of results and multiply by 100. In this case, the frequency of the first row is 1 and the total number of results is 10. The percentage would then be 10.0.
      10.0. (1 ÷ 10) X 100 = 10.0
    - The final column is **Cumulative percentage**. In this column, divide the cumulative frequency by the total number of results and then to make a percentage, multiply by 100. Note that the last number in this column should always equal 100.0. In this example, the cumulative frequency is 1 and the total number of results is 10, therefore the cumulative percentage of the first row is 10.0.
      10.0. (1 ÷ 10) X 100 = 10.0

The cumulative frequency distribution table should look like this:

Table 4.3.2
Ages of participants at a chess tournament

Lower Value	Upper Value	Frequency (f)	Cumulative frequency	Percentage	Cumulative percentage
35	44	1	1	10.0	10.0
45	54	2	3	20.0	30.0
55	64	2	5	20.0	50.0
65	74	2	7	20.0	70.0
75	84	2	9	20.0	90.0
85	94	1	10	10.0	100.0

Class intervals

If a variable takes a large number of values, then it is easier to present and handle the data by grouping the values into class intervals. Continuous variables are more likely to be presented in class intervals, while discrete variables can be grouped into class intervals or not.

To illustrate, suppose we set out age ranges for a study of young people, while allowing for the possibility that some older people may also fall into the scope of our study.

The **frequency** of a class interval is the number of observations that occur in a particular predefined interval. So, for example, if 20 people aged 5 to 9 appear in our study's data, the frequency for the 5–9 interval is 20.

The **endpoints** of a class interval are the lowest and highest values that a variable can take. So, the intervals in our study are 0 to 4 years, 5 to 9 years, 10 to 14 years, 15 to 19 years, 20 to 24 years, and 25 years and over. The endpoints of the first interval are 0 and 4 if the variable is discrete, and 0 and 4.999 if the variable is continuous. The endpoints of the other class intervals would be determined in the same way.

Class interval width is the difference between the lower endpoint of an interval and the lower endpoint of the next interval. Thus, if our study's continuous intervals are 0 to 4, 5 to 9, etc., the width of the first five intervals is 5, and the last interval is open, since no higher endpoint is assigned to it. The intervals could also be written as 0 to less than 5, 5 to less than 10, 10 to less than 15, 15 to less than 20, 20 to less than 25, and 25 and over.

Rules for data sets that contain a large number of observations

In summary, follow these basic rules when constructing a frequency distribution table for a data set that contains a large number of observations:

- find the lowest and highest values of the variables
- decide on the width of the class intervals
- include all possible values of the variable.

In deciding on the width of the class intervals, you will have to find a compromise between having intervals short enough so that not all of the observations fall in the same interval, but long enough so that you do not end up with only one observation per interval.

It is also important to make sure that the class intervals are mutually exclusive and collectively exhaustive.

Example 3 – Constructing a frequency distribution table for large numbers of observations

Thirty AA batteries were tested to determine how long they would last. The results, to the nearest minute, were recorded as follows:

423, 369, 387, 411, 393, 394, 371, 377, 389, 409, 392, 408, 431, 401, 363, 391, 405, 382, 400, 381, 399, 415, 428, 422, 396, 372, 410, 419, 386, 390

Use the steps in Example 1 and the above rules to help you construct a frequency distribution table.

Answer

The lowest value is 363 and the highest is 431.

Using the given data and a class interval of 10, the interval for the first class is 360 to 369 and includes 363 (the lowest value). Remember, there should always be enough class intervals so that the highest value is included.

The completed frequency distribution table should look like this:

Table 4.3.3
Life of AA batteries, in minutes

Battery life, minutes (x)	Frequency (f)
360–369	2
370–379	3
380–389	5
390–399	7
400–409	5
410–419	4
420–429	3
430–439	1
Total	**30**

Example 4 – Constructing relative frequency and percentage frequency tables

An analyst studying the data from example 3 might want to know not only how long batteries last, but also what proportion of the batteries falls into each class interval of battery life.

This **relative frequency** of a particular observation or class interval is found by dividing the frequency (f) by the number of observations (n): that is, (f ÷ n). Thus:

Relative frequency = frequency ÷ number of observations

The **percentage frequency** is found by multiplying each relative frequency value by 100. Thus:

Percentage frequency = relative frequency X 100 = f ÷ n X 100

Use the data from Example 3 to make a table giving the relative frequency and percentage frequency of each interval of battery life.

Here is what that table looks like:

Table 4.3.4
Life of AA batteries, in minutes

Battery life, minutes (x)	Frequency (f)	Relative frequency	Percent frequency
360–369	2	0.07	7
370–379	3	0.1	10
380–389	5	0.17	17
390–399	7	0.23	23
400–409	5	0.17	17
410–419	4	0.13	13
420–429	3	0.1	10
430–439	1	0.03	3
Total	**30**	**1**	**100**

An analyst of these data could now say that:

- 7% of AA batteries have a life of from 360 minutes up to but less than 370 minutes, and that
- the probability of any randomly selected AA battery having a life in this range is approximately 0.07.

Example 5 – Visualization of the cumulative relative frequency distribution

As previously shown for example 2, **cumulative frequency** is used to determine the number of observations that lie below a particular value in a data set. The cumulative frequency is calculated by adding each frequency from a frequency distribution table to the sum of its predecessors. The last value will always be equal to the total for all observations, since all frequencies will already have been added to the previous total. Let's look at another example of how to calculate the cumulative frequency.

The daily number of rock climbers in Lake Louise, Alberta was recorded over a 30-day period. The results are as follows:

31, 49, 19, 62, 24, 45, 23, 51, 55, 60, 40, 35 54, 26, 57, 37, 43, 65, 18, 41, 50, 56, 4, 54, 39, 52, 35, 51, 63, 42.

The number of rock climbers ranges from 4 to 65. In order to create a frequency table, the data are best grouped in class intervals of 10. Each interval can be one row in the frequency table. The **Frequency** column lists the number of observations found within a class interval. For example, there are only two values in the interval from 10 to 20, then its frequency is 2 in the table accordingly.

Use the **Frequency** column to calculate cumulative frequency.

1. First, add the number from the **Frequency** column to its predecessor. For example, in the first row, we have only one observation and no predecessors. The cumulative frequency is one.
   1 + 0 = 1
2. However, in the second row, there are two observations. Add these two to the previous cumulative frequency (one), and the result is three.
   1 + 2 = 3
3. Record the results in the **Cumulative frequency** column.

The other entries in the table can be calculated similarly. Results are presented in the table 4.3.5.

Table 4.3.5
Frequency and cumulative frequency of daily number of rock climbers recorded in Lake Louise, Alberta, 30-day period

Number of rock climbers	Frequency (f)	Cumulative frequency
<10	1	1
10 to <20	2	1 + 2 = 3
20 to <30	3	3 + 3 = 6
30 to <40	5	6 + 5 = 11
40 to <50	6	11 + 6 = 17
50 to <60	9	17 + 9 = 26
>= 60	4	26 + 4 = 30

Cumulative relative frequency is another way of expressing frequency distribution. It is obtained by calculating the percentage of the cumulative frequency within each interval.

Cumulative percentage is calculated by dividing the cumulative frequency by the total number of observations (**n**), then multiplying it by 100 (the last value will always be equal to 100%). Thus,

cumulative relative frequency = (cumulative frequency ÷ n) x 100

The fourth column in the table 4.3.6 shows the calculation of the cumulative relative frequency of the daily number of rock climbers recorded in Lake Louise.

Table 4.3.6
Cumulative relative frequency of daily number of rock climbers recorded in Lake Louise, Alberta, 30-day period

Number of rock climbers	Frequency (f)	Cumulative frequency	Cumulative relative frequency (%)
<10	1	1	1 ÷ 30 x 100 = 3
10 to <20	2	1 + 2 = 3	3 ÷ 30 x 100 = 10
20 to <30	3	3 + 3 = 6	6 ÷ 30 x 100 = 20
30 to <40	5	6 + 5 = 11	11 ÷ 30 x 100 = 37
40 to <50	6	11 + 6 = 17	17 ÷ 30 x 100 = 57
50 to <60	9	17 + 9 = 26	26 ÷ 30 x 100 = 87
>= 60	4	26 + 4 = 30	30 ÷ 30 x 100 = 100

The cumulative relative frequency distribution can be visualized with a bar chart or a line chart, like in chart 4.3.1 below. The value on the horizontal axis is the upper bound of the class interval.

Chart 4.3.1
Cumulative relative frequency of the daily number of rock climbers in Lake Louise, Alberta, during a 30-day period

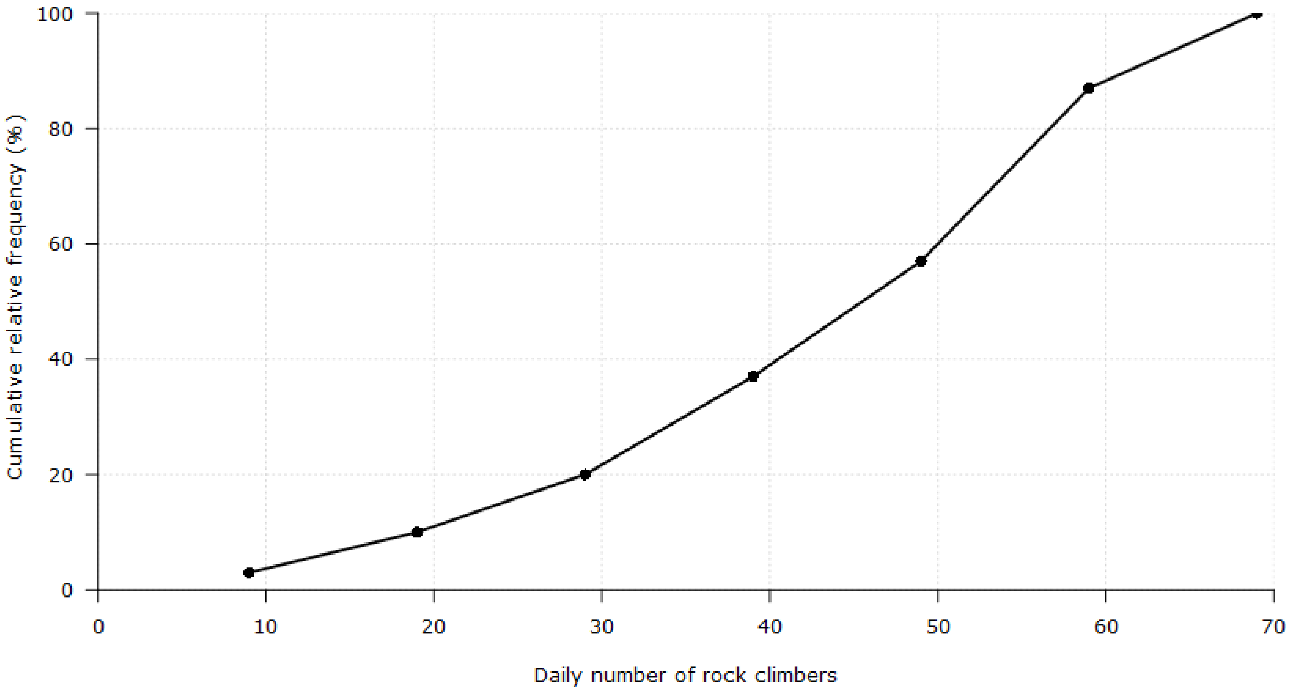


Chart 4.3.1 shows that for the majority of days (57%) in the period, the number of rock climbers was lower or equal to 49.

Frequency distribution can be visualized using:

- a pie chart (nominal variable),
- a bar chart (nominal or ordinal variable),
- a line chart (ordinal or discrete variable),
- or a histogram (continuous variable).

These types of charts will be presented in the section 5 on data visualization. But first, we will look at other methods to summarize data using measures of central tendency and dispersion.

4.4 Measures of central tendency

The best way to summarize a data set with a single value is to find the most representative value, the one that indicates where the centre of the distribution is. This is called the central tendency. The three most commonly used measures of central tendency are

- The **arithmetic mean**, which is the sum of all values divided by the number of values,
- The **median**, which is the middle value when all values are arranged in increasing order,
- The **mode**, which is the most typical value, the one that appears the most often in the data set.

In the following sections, the way to calculate these three measures will be explained with the help of examples.

4.4.1 Calculating the mean

The mean can be calculated only for numeric variables, no matter if they are discrete or continuous. It's obtained by simply dividing the sum of all values in a data set by the number of values. The calculation can be done from raw data or for data aggregated in a frequency table. Here are a few examples of calculation.

Example 1 – Soccer tournament at Mount Rival

Mount Rival hosts a soccer tournament each year. This season, in 10 games, the lead scorer for the home team scored 7, 5, 0, 7, 8, 5, 5, 4, 1 and 5 goals. What is the mean score of this player?

The sum of all values is 47 and there are 10 values. Therefore, the mean is 47 ÷ 10 = 4.7 goals per game.

Example 2 – Traffic fatalities

The following table lists the number of people killed in traffic accidents over a 10-year period. During this period, what was the average number of people having lost life every year? How many people died each day on average in traffic accidents?

Table 4.4.1.1
Number of fatalities in traffic accidents

Year	Deaths
2009	623
2010	583
2011	959
2012	1,037
2013	960
2014	797
2015	663
2016	652
2017	560
2018	619
Total	**7,453**

The total number of deaths is presented in the table (7,453). To get the yearly average, the number of deaths is divided by 10 which gives 745.3 deaths per year. To get the daily average, the yearly average is divided by 365 which gives approximately 2 deaths per day.

For a larger data set, it can be easier to summarize data in a frequency table before calculating the mean. In this case, you need to weight each possible value by the frequency of the value to calculate the total.

Example 3 – Soccer tournament at Mount Rival

Let's go back to Mount Rival soccer tournament. Suppose that five teams were competing, each of them including 10 players for a total of 50 players. The number of goals scored by each player was compiled and results were summarized in the frequency table below. For example, we can see that eight players scored only one goal during the tournament. What is the average number of goals scored by the players during the tournament?

Table 4.4.1.2
Number of players by the number of goals scored

Number of goals scored	Number of players
0	2
1	8
2	14
3	12
4	8
5	4
6	2

0 true zero or a value rounded to zero

You first need to calculate the total number of goals scored. To do that, you take each observed value of the number of goals scored, which are values 0 to 6, and you multiply each value by the number of players:

$0 \times 2 + 1 \times 8 + 2 \times 14 + 3 \times 12 + 4 \times 8 + 5 \times 4 + 6 \times 2 = 136$

Since there are 50 players, the average is $136 \div 50 = 2.72$ goals per player.

4.4.2 Calculating the median

The median is the value in the middle of a data set, meaning that 50% of data points have a value smaller or equal to the median and 50% of data points have a value higher or equal to the median. For a small data set, you first count the number of data points (n) and arrange the data points in increasing order. If the number of data points is uneven, you add 1 to the number of points and divide the results by 2 to get the rank of the data point whose value is the median. The rank is the position of the data point after the data set has been arranged in increasing order: the smallest value is rank 1, the second-smallest value is rank 2, etc.

Example 1 – Median time in 200 metres of a top running athlete

Imagine that a top running athlete in a typical 200-metre training session runs in the following times: 26.1 seconds, 25.6 seconds, 25.7 seconds, 25.2 seconds, 25.0 seconds, 27.8 seconds and 24.1 seconds. How would you calculate his median time?

Let's start with arranging the values in increasing order:

Table 4.4.2.1
Rank associated with each value of 200-meter running times

Rank	Times (in seconds)
1	24.1
2	25.0
3	25.2
4	25.6
5	25.7
6	26.1
7	27.8

There are n = 7 data points, which is an uneven number. The median will be the value of the data points of rank

(n + 1) ÷ 2 = (7 + 1) ÷ 2 = 4.

The median time is 25.6 seconds.

If the number of data points is even, the median will be the average of the data point of rank n ÷ 2 and the data point of rank (n ÷ 2) + 1.

Example 2 – Median time in 200 metres of a top running athlete (Part 2)

Now suppose that the athlete runs his eighth 200-metre run with a time of 24.7 seconds. What is his median time now?

Table 4.4.2.2
Rank associated with each value of 200-meter running times, updated

Rank	Times (in seconds)
1	24.1
2	24.7
3	25.0
4	25.2
5	25.6
6	25.7
7	26.1
8	27.8

There are now n = 8 data points, an even number. The median is the mean between the data point of rank

n ÷ 2 = 8 ÷ 2 = 4

and the data point of rank

(n ÷ 2) + 1 = (8 ÷ 2) +1 = 5

Therefore, the median time is (25.2 + 25.6) ÷ 2 = 25.4 seconds.

For larger data sets, the cumulative relative frequency distribution can be helpful to identify the median. The median is the smallest value for which the cumulative relative frequency is at least 50%. However, when possible it's best to use the basic statistical function available in a spreadsheet or statistical software application because the results will then be more reliable.

Example 3 – Median size of households of the students in the class

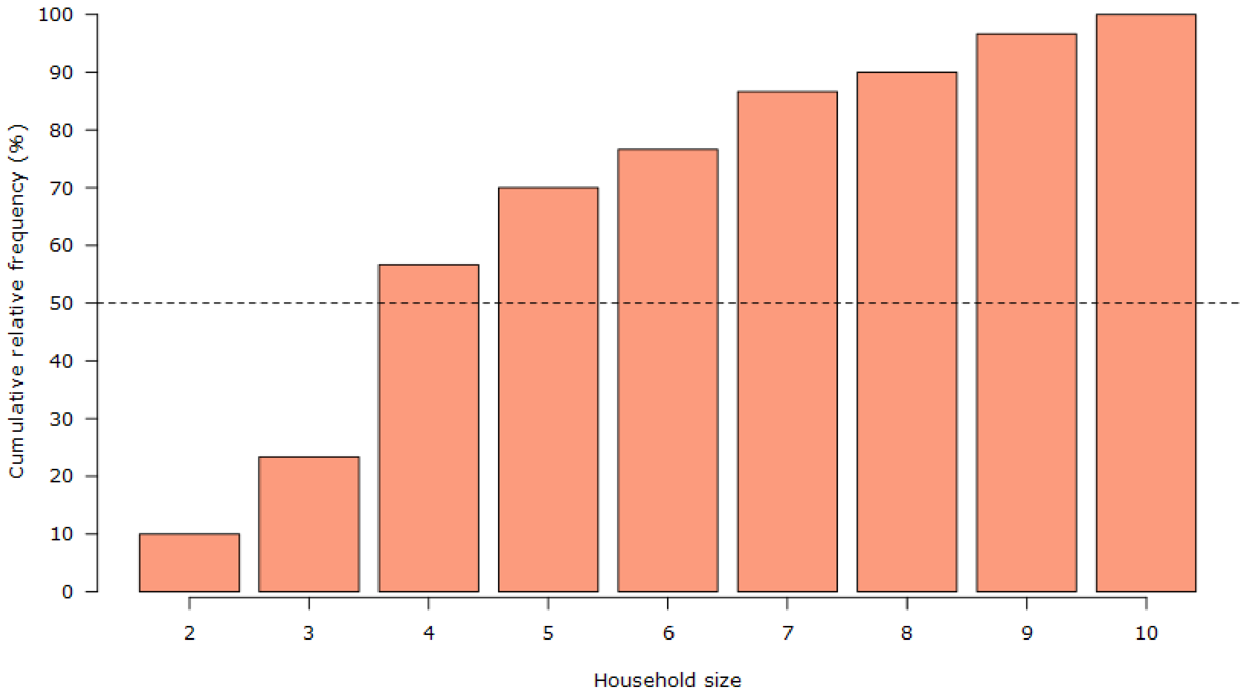
IImagine you ask the 30 students of your class how many people there are in their households. You summarize the data you collected in a frequency table, in which you include the relative frequencies and the cumulative relative frequencies.

Table 4.4.2.3
Frequency table of household sizes of the students

Household size	Frequency (number of students)	Relative frequency (%)	Cumulative frequency (number of students)	Cumulative relative frequency (%)
2	3	10.0	3	10.0
3	4	13.3	7	23.3
4	10	33.3	17	56.7
5	4	13.3	21	70.0
6	2	6.7	23	76.7
7	3	10.0	26	86.7
8	1	3.3	27	90.0
9	2	6.7	29	96.7
10	1	3.3	30	100.0

You can see that 10% of students (3 students) live in a household of size 2, 23% of students (7 students) live in a household of size 3 or less and 57% of students (17 students) live in a household of size 4 or less. The median will be equal to 4 because it's the smallest value for which the cumulative relative frequency is higher than 50%. This is even more obvious if you visualize the cumulative relative frequency on a bar chart like on chart 4.4.2.1. The dotted line indicates the cumulative relative frequency of 50%.

Chart 4.4.2.1
Cumulative relative frequency of the household size of students in the class



The mean is the total number of people in the households of the students:

2 × 3 + 3 × 4 + 4 × 10 + 5 × 4 + 6 × 2 + 7 × 3 + 8 × 1 + 9 × 2 + 10 × 1 = 147

divided by the number of students, which is 30. The result is 147 ÷ 30 = 4.9 people per household.

In this example, the median (4) is lower than the mean (4.9).

The advantage of using the median instead of the mean is that the median is more robust, which means that an extreme value added to one extremity of the distribution don't have an impact on the median as big as the impact on the mean. Therefore, it is important to check if the data set includes extreme values before choosing a measure of central tendency. This will be illustrated by the next example.

Example 4 – Median size of households of the students in the class (Part 2)

A new student recently joined your class. You decide to ask him what the size of his household is in order to update your results. He replies to you that he lives in a large multi-generational house that includes 18 people!

Once updated, the mean is (147 + 18) ÷ 31 = 5.3 people per household. Just adding one new student increased the mean by 0.4 (5.3 – 4.9). The median is the same after the update. There are now 7 ÷ 31 = 22.6% of students in a household of size 3 or less, and 17 ÷ 31 = 54.8% of students that live in a household of size 4 or less. The value 4 is still the smallest value with a cumulative relative frequency of at least 50%.

4.4.3 Calculating the mode

When it's unique, the mode is the value that appears the most often in a data set and it can be used as a measure of central tendency, like the median and mean. But sometimes, there is no mode or there is more than one mode.

There is no mode when all observed values appear the same number of times in a data set. There is more than one mode when the highest frequency was observed for more than one value in a data set. In both of these cases, the mode can't be used to locate the centre of the distribution.

The mode can be used to summarize categorical variables, while the mean and median can be calculated only for numeric variables. This is the main advantage of the mode as a measure of central tendency. It's also useful for discrete variables and for continuous variables when they are expressed as intervals.

Here are some examples of calculation of the mode for discrete variables.

Example 1 – Number of points during a hockey tournament

During a hockey tournament, Audrey scored 7, 5, 0, 7, 8, 5, 5, 4, 1 and 5 points in 10 games. After summarizing the data in a frequency table, you can easily see that the mode is 5 because this value appears the most often in the data set (4 times). The mode can be considered a measure of central tendency for this data set because it's unique.

Table 4.4.3.1
Number of games by the number of points scored

Number of points scored	Frequency (number of games)
0	1
1	1
4	1
5	4
7	2
8	1

0 true zero or a value rounded to zero

Example 2 – Number of points in 12 basketball games

During Marco's 12-game basketball season, he scored 14, 14, 15, 16, 14, 16, 16, 18, 14, 16, 16 and 14 points. After summarizing the data in a frequency table, you can see that there are two modes in this data set: 14 and 16. Both values appear 5 times in the data set and 5 is the highest frequency observed. The mode can't be used a measure of central tendency because there is more than one mode. It's a bimodal distribution.

Table 4.4.3.2
Number of games by the number of points scored

Number of points scored	Frequency (number of games)
14	5
15	1
16	5
18	1

Example 3 – Number of touchdowns scored during football season

The following data set represents the number of touchdowns scored by Jerome in his high-school football season: 0, 0, 1, 0, 0, 2, 3, 1, 0, 1, 2, 3, 1, 0. Let's compare the mean, median and mode.

The sum of all values is 14 and there are 14 data points. This gives a mean of 1. Because the number of values is even, the median is average between the data point of rank 7 and the data point of rank 8, after arranging the data set in increasing order.

Table 4.4.3.3
Rank associated with each value of the number of touchdowns during football season

Rank	Number of touchdowns
1	0
2	0
3	0
4	0
5	0
6	1
7	1
8	1
9	1
10	1
11	2
12	2
13	3
14	3

Therefore, the median is equal to 1. Once the data has been summarized in a frequency table, you can see that the mode is 0 because it is the value that appears the most often (6 times).

Table 4.4.3.4
Number of games by the number of touchdowns

Number of touchdowns	Frequency
0	6
1	4
2	2
3	2

0 true zero or a value rounded to zero

In summary, in this example, the mean is 1, the median is 1 and the mode is 0.

The mode is not used as much for continuous variables because with this type of variable, it is likely that no value will appear more than once. For example, if you ask 20 people their personal income in the previous year, it's possible that many will have amounts of income that are very close, but that you will never get exactly the same value for two people. In such case, it is useful to group the values in mutually exclusive intervals and to visualize the results with a histogram to identify the modal-class interval.

Example 4 – Height of people in the arena during a basketball game

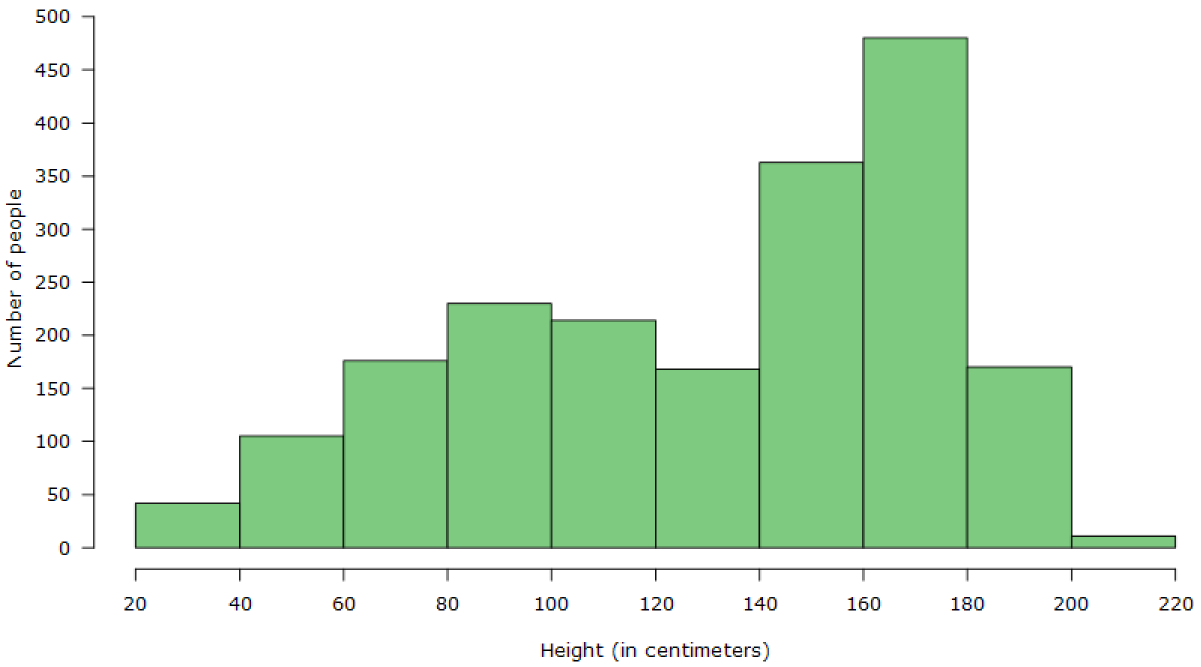
We are interested in the height of the people present in the arena during a basketball game. Table 4.4.3.5 presents the number of people for 20-centimetre intervals of height.

Table 4.4.3.5
Number of people by height intervals

Height (in centimetres)	Frequency (number of people)
20 to 39	42
40 to 59	105
60 to 79	176
80 to 99	230
100 to 119	214
120 to 139	168
140 to 159	363
160 to 179	480
180 to 200	170
200 to 219	11

Chart 4.4.3.1 shows this data set as a histogram.

Chart 4.4.3.1
Histogram of the height of people at the basketball match



For categorical or discrete variables, multiple modes are values that reach the same frequency: the highest one observed. For continuous variables, all peaks of the distribution can be considered modes even if they don't have the same frequency. The distribution for this example is bimodal, with a major mode corresponding to the modal-class interval 160 to 179 centimetres and a minor mode corresponding to the modal-class interval 80 to 99 centimetres. The modal class shouldn't be used as a measure of central tendency, but finding two modes gives us an indication that there could be two distinct groups in the data that should be analyzed separately.

4.5 Measures of dispersion

Measures of central tendency aim to identify the most representative value of a data set, that is, the centre of a distribution. To better describe the data, it is also good to have a measure of the spread of the data around the centre of the distribution. This measure is called a measure of dispersion. The most commonly used measures of dispersion are

- The **range**, which is the difference between the highest value and the smallest value;
- The **interquartile range**, which is the range of the 50% of data that is central to the distribution;
- The **variance**, which is the mean squared distance between each point and the centre of the distribution;
- The **standard deviation**, which is the square root of variance.

The following sections explain how to calculate these measures using examples. Measures of dispersion are applicable to numeric variables only.

4.5.1 Calculating the range and interquartile range

To calculate the range, you need to find the largest observed value of a variable (the maximum) and subtract the smallest observed value (the minimum). The range only takes into account these two values and ignore the data points between the two extremities of the distribution. It's used as a supplement to other measures, but it is rarely used as the sole measure of dispersion because it's sensitive to extreme values.

The interquartile range and semi-interquartile range give a better idea of the dispersion of data. To calculate these two measures, you need to know the values of the lower and upper quartiles. The lower quartile, or first quartile (Q1), is the value under which 25% of data points are found when they are arranged in increasing order. The upper quartile, or third quartile (Q3), is the value under which 75% of data points are found when arranged in increasing order. The median is considered the second quartile (Q2). The interquartile range is the difference between upper and lower quartiles. The semi-interquartile range is half the interquartile range.

When the data set is small, it is simple to identify the values of quartiles. Let's look at an example.

Example 1 – Range and interquartile range of a data set

Find the quartiles of this data set: 6, 47, 49, 15, 43, 41, 7, 39, 43, 41, 36.

You first need to arrange the data points in increasing order. As you do so, you can give them a rank to indicate their position in the data set. Rank 1 is the data point with the smallest value, rank 2 is the data point with the second-lowest value, etc.

Table 4.5.1.1
Rank of data points

Rank	Value
1	6
2	7
3	15
4	36
5	39
6	41
7	41
8	43
9	43
10	47
11	49

Then you need to find the rank of the median to split the data set in two. As we have seen in the section on the median, if the number of data points is an uneven value, the rank of the median will be

(n + 1) ÷ 2 = (11 + 1) ÷ 2 = 6

The rank of the median is 6, which means there are five points on each side.

Then you need to split the lower half of the data in two again to find the lower quartile. The lower quartile will be the point of rank (5 + 1) ÷ 2 = 3. The result is Q1 = 15. The second half must also be split in two to find the value of the upper quartile. The rank of the upper quartile will be 6 + 3 = 9. So Q3 = 43.

Once you have the quartiles, you can easily measure the spread. The interquartile range will be Q3 - Q1, which gives 28 (43-15). The semi-interquartile range is 14 (28 ÷ 2) and the range is 43 (49-6).

For larger data sets, you can use the cumulative relative frequency distribution to help identify the quartiles or, even better, the basic statistics functions available in a spreadsheet or statistical software that give results more easily.

What happens when the data set includes a data point whose value is considered extreme compared to the rest of the distribution?

Example 2 – Range and interquartile range in presence of an extreme value

Find the range and interquartile range of the data set of example 1, to which a data point of value 75 was added.

The range would now be 69 (75-6). The median would be the mean of the values of the data point of rank 12 ÷ 2 = 6 and the data point of rank (12 ÷ 2) + 1 = 7. Because it falls between ranks 6 and 7, there are six data points on each side of the median. The lower quartile is the mean of the values of the data point of rank 6 ÷ 2 = 3 and the data points of rank (6 ÷ 2) + 1 = 4. The result is (15 + 36) ÷ 2 = 25.5. The upper quartile is the mean of the values of data point of rank 6 + 3 = 9 and the data point of rank 6 + 4 = 10, which is (43 + 47) ÷ 2 = 45. The interquartile range is 45 - 25.5 = 19.5.

In summary, the range went from 43 to 69, an increase of 26 compared to example 1, just because of a single extreme value. The more robust interquartile range went from 28 to 19.5, a decrease of only 8.5.

The second example demonstrated that the interquartile range is more robust than the range when the data set includes a value considered extreme. It's not a perfect measure, though. In this example, we might have expected that when adding an extreme value, the measure of dispersion would increase, but the opposite happened because there was a great difference between the values of data points of ranks 3 and 4.

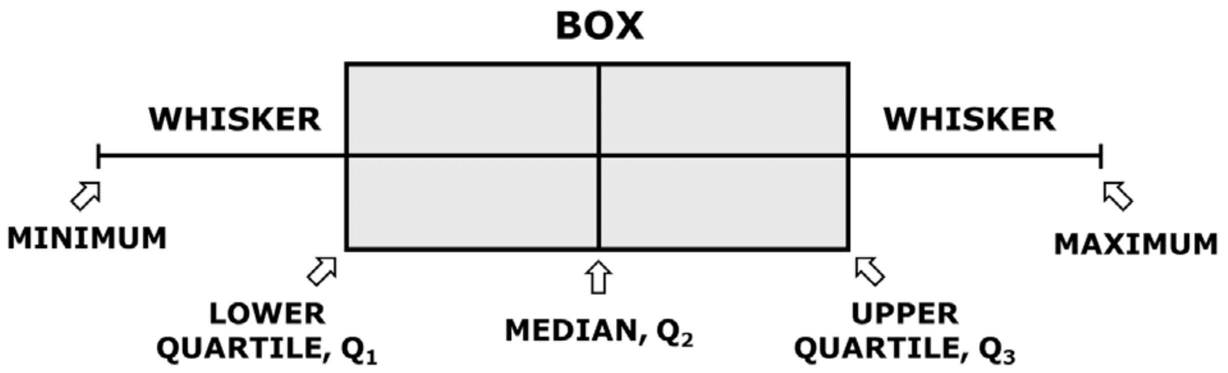
The five-value series formed by the minimum, the three quartiles and the maximum is often referred to as "the five-number summary." It is a well-known manner to summarize data sets. In the following section on box and whisker plot, we will see a useful method to visualize this five-number summary.

4.5.2 Visualizing the box and whisker plot

The box and whisker plot, sometimes simply called the box plot, is a type of graph that help visualize the five-number summary. It doesn't show the distribution in as much detail as histogram does, but it's especially useful for indicating whether a distribution is skewed and whether there are potential unusual observations (outliers) in the data set. A box plot is ideal for comparing distributions because the centre, spread and overall range are immediately apparent.

Figure 4.5.2.1 shows how to build the box and whisker plot from the five-number summary.

Figure 4.5.2.1 Building a box and whisker plot



In a box and whisker plot:

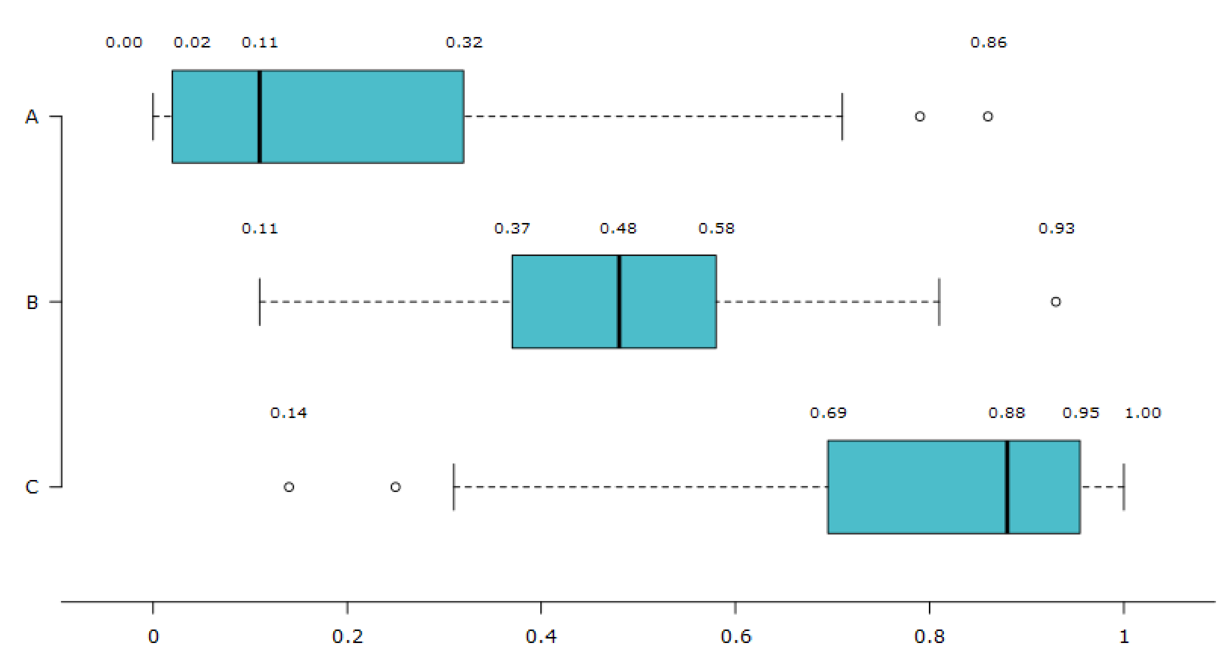
- The left and right sides of the box are the lower and upper quartiles. The box covers the interquartile interval, where 50% of the data is found.
- The vertical line that split the box in two is the median. Sometimes, the mean is also indicated by a dot or a cross on the box plot.
- The whiskers are the two lines outside the box, that go from the minimum to the lower quartile (the start of the box) and then from the upper quartile (the end of the box) to the maximum.
- The graph is usually presented with an axis that indicates the values (not shown on figure 4.5.2.1).
- The box and whisker plot can be presented horizontally, like in figure 4.5.2.1, or vertically.

A variation of the box and whisker plot restricts the length of the whiskers to a maximum of 1.5 times the interquartile range. That is, the whisker reaches the value that is the furthest from the centre while still being inside a distance of 1.5 times the interquartile range from the lower or upper quartile. Data points that are outside this interval are represented as points on the graph and considered potential outliers.

Example 1 – Comparison of three box and whisker plots

The three box and whisker plots of chart 4.5.2.1 have been created using R software. What can you say about the three distributions?

Chart 4.5.2.1
Box and whisker plots and five-number summaries of distributions A, B and C



- The centre of distribution A is the lowest of the three distributions (median is 0.11). The distribution is positively skewed, because the whisker and half-box are longer on the right side of the median than on the left side.
- Distribution B is approximately symmetric, because both half-boxes are almost the same length (0.11 on the left side and 0.10 on the right side). It's the most concentrated distribution because the interquartile range is 0.21, compared to 0.30 for distribution A and 0.26 for distribution C.
- The centre of distribution C is the highest of the three distributions (median is 0.88). The distribution C is negatively skewed because the whisker and half-box are longer on the left side of the median than on the right side.

All three distributions include potential outliers. Let's take distribution A, for example. The interquartile range is Q3 - Q1 = 0.32 – 0.02 = 0.30. According to the definition used by the function in R software, all values higher than Q3 + 1.5 x (Q3 - Q1) = 0.32 + 1.5 x 0.30 = 0.77 are outside the right whisker and indicated by a circle. There are two potential outliers in distribution A.

4.5.3 Calculating the variance and standard deviation

Unlike range and interquartile range, variance is a measure of dispersion that takes into account the spread of all data points in a data set. It's the measure of dispersion the most often used, along with the standard deviation, which is simply the square root of the variance. The variance is mean squared difference between each data point and the centre of the distribution measured by the mean.

Example 1 – Calculation of variance and standard deviation

Let's calculate the variance of the follow data set: 2, 7, 3, 12, 9.

The first step is to calculate the mean. The sum is 33 and there are 5 data points. Therefore, the mean is 33 ÷ 5 = 6.6. Then you take each value in data set, subtract the mean and square the difference. For instance, for the first value:

$(2 - 6.6)^2 = 21.16$

The squared differences for all values are added:

21.16 + 0.16 + 12.96 + 29.16 + 5.76 = 69.20

The sum is then divided by the number of data points:

69.20 ÷5 = 13.84

The variance is 13.84. To get the standard deviation, you calculate the square root of the variance, which is 3.72.

Standard deviation is useful when comparing the spread of two separate data sets that have approximately the same mean. The data set with the smaller standard deviation has a narrower spread of measurements around the mean and therefore usually has comparatively fewer high or low values. An item selected at random from a data set whose standard deviation is low has a better chance of being close to the mean than an item from a data set whose standard deviation is higher. However, standard deviation is affected by extreme values. A single extreme value can have a big impact on the standard deviation.

Standard deviation might be difficult to interpret in terms of how large it has to be when considering the data to be widely dispersed. The magnitude of the mean value of the dataset affects the interpretation of its standard deviation. When you are measuring something that is in the scale of millions, having measures that are close to the mean value doesn't have the same meaning as when you are measuring something that is in the scale of hundreds. For example, a measure of two large companies with a difference of $10,000 in annual revenues is considered pretty close, while the measure of two individuals with a weight difference of 30 kilograms is considered far apart. This is why, in most situations, it is helpful to assess the size of the standard deviation relative to its mean.

Remember the following properties when you are using the standard deviation:

- Standard deviation is sensitive to extreme values. A single very extreme value can increase the standard deviation and misrepresent the dispersion.
- For two data sets with the same mean, the one with the larger standard deviation is the one in which the data is more spread out from the center.
- Standard deviation is equal to 0 if all values are equal (because all values are then equal to the mean).

The reason why standard deviation is so popular as a measure of dispersion is its relation with the normal distribution which describes many natural phenomena and whose mathematical properties are interesting in the case of large data sets. When a variable follows a normal distribution, the histogram is bell-shaped and symmetric, and the best measures of central tendency and dispersion are the mean and the standard deviation. It's a very useful probability distribution and relatively easy to use. Confidence intervals are often based on the standard normal distribution.

However, when:

- the data set is small,
- the distribution is asymmetric, or
- the data set includes extreme values

it's better to use the interquartile range.

4.6 Exercises

1. Indicate whether each of the following variables is discrete or continuous:

   a. the time it takes for you to get to school
   b. the number of Canadian couples who were married last year
   c. the number of goals scored by a women's hockey team
   d. the speed of a bicycle
   e. your age
   f. the number of subjects your school offered last year
   g. the length of time of a telephone call
   h. the annual income of an individual
   i. the distance between your house and school
   j. the number of pages in a dictionary

2. A local convenience store owner records how many customers enter the store each day over a 25-day period. The results are as follows:

20, 21, 23, 21, 26, 24, 20, 24, 25, 22, 22, 23, 21, 24, 21, 26, 24, 22, 21, 23, 25, 22, 21, 24, 21

   a. Present these data in a frequency distribution table.
   b. Which result occurs most frequently?
   c. Set up a frequency distribution table including columns for the relative frequency and percentage frequency of the data.

3. Forty students took a math test marked out of 10 points. Their results were as follows:

9, 10, 7, 8, 9, 6, 5, 9, 4, 7, 1, 7, 2, 7, 8, 5, 4, 3, 10, 7, 3, 7, 8, 6, 9, 7, 4, 2, 3, 9, 4, 3, 7, 5, 5, 2, 7, 9, 7, 1

   a. Prepare a frequency table of the scores.
   b. Using the frequency table, calculate the mean, median and mode.
   c. Interpret these results.

4. The following table outlines hypothetical numbers of new hires at a large organization over a ten-year period:

Table 4.6.1
Hypothetical number of new hires

Year	Number of new hires
1	266
2	231
3	223
4	262
5	260
6	230
7	191
8	182
9	165
10	153

a. Find the range.

b. Calculate the interquartile range.

c. Calculate the five-number summary.

d. Draw a box and whisker plot for this data.

4.7 Answers

1. a. This is a continuous variable.
   b. This is a discrete variable.
   c. This is a discrete variable.
   d. This is a continuous variable.
   e. The exact age is a continuous variable, but age is often rounded down to the closest integer. In this case, it would be a discrete variable.
   f. This is a discrete variable.
   g. This is a continuous variable.
   h. This is a continuous variable.
   i. This is a continuous variable.
   j. This is a discrete variable.

2. a.

Table 4.7.1
Answer for 2a

Number of customers (x)	Frequency (f)
20	2
21	7
22	4
23	3
24	5
25	2
26	2
Total	**25**

b. The observation that occurs the most frequently is 21.

c.

Table 4.7.2
Answer for 2c

Number of customers (x)	Frequency (f)	Relative frequency	Percentage frequency
20	2	0.08	8
21	7	0.28	28
22	4	0.16	16
23	3	0.12	12
24	5	0.20	20
25	2	0.08	8
26	2	0.08	8
Total	25	1.00	100

3. a.

Table 4.7.3
Answer for 3a

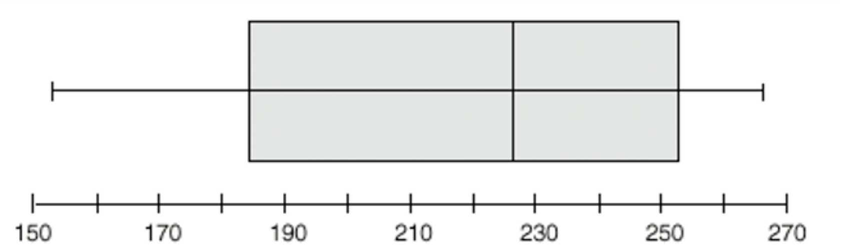
Score (x)	Frequency (f)
1	2
2	3
3	4
4	4
5	4
6	2
7	10
8	3
9	6
10	2
Total	**40**

b. mean = 5.9, median = 7, mode = 7

c. The median is higher than the mean because most of the observations have high values. The mean is influenced by the lower scores. The mode is equal to the median.

4. a. 113
b. 78
c. 153, 182, 226.5, 260, 266

Chart 4.7.1
Box and whisker plot for the hypothetical number of new hires



5 Data Visualization

Graphs and charts are effective visual tools because they present information quickly and easily. It is not surprising then, that graphs are commonly used by print and electronic media. Sometimes, data can be better understood when presented by a graph than by a table because the graph can reveal a **trend or comparison.**

Students also find that graphs are easy to use because graphs are made up of lines, dots and blocks—all geometric forms that are simple and quick for students to draw. In the world of statistics, graphs display the relationship between variables or show the value spread of a given variable or phenomenon.

This section aims to describe the graphs the most often used to visualize data. There are many other graphs that can be used in different contexts, such as the heat map, the tree map, the bubble chart, the area chart, the radar chart as well as the box and whisker plot that has been presented in a previous section.

5.1 Using graphs

Knowing how to convey information graphically is important in presenting statistics. The following is a list of general rules to keep in mind when preparing graphs.

A good graph:

- accurately shows the facts,
- grabs the reader's attention,
- complements or demonstrates arguments presented in the text,
- has a title and labels,
- is simple and uncluttered,
- shows data without altering the message of the data,
- clearly shows any trends or differences in the data,
- is visually accurate (i.e. if one chart value is 15 and another 30, then 30 should appear to be twice the size of 15).

Why use graphs to present data?

Because they…

- are quick and direct,
- highlight the most important facts,
- facilitate understanding of the data,
- can convince readers,
- can be easily remembered.

There are many different types of graphs that can be used to convey information, including:

- bar charts
- pictographs
- pie charts
- line charts
- scatterplots
- histograms

Knowing what type of graph or chart to use with what type of information is crucial. Depending on the nature of the variables, some types are more appropriate than others. For example, categorical variables like school subjects are best displayed in a bar chart or pie chart while continuous variables such as height are illustrated by a line chart or histogram.

Graphs: four guidelines

If you have decided that using a graph is the best method to relay your message, then there are four guidelines to remember:

1. **Define your target audience**

Ask yourself the following questions to help you understand more about your audience and what their needs are:

- Who is your target audience?
- What do they know about the issue?
- What do they expect to see?
- What do they want to know?
- How will they use the information?

2. **Determine the message(s) to be transmitted**

Ask yourself the following questions to figure out what your message is and why it is important:

- What do the data show?
- Is there more than one main message?
- What aspect of the message(s) should be highlighted?
- Can all the messages be displayed in the same graph or chart?

3. **Use appropriate terms to describe your graph**

Consider the following appropriate terms when labelling the graph or describing features of it in accompanying text:

Table 5.1.1
Terms to describe graphs

If your graph…	Use the following terms…
describes components	share of, percent of the, smallest, the majority of
compares items	ranking, larger than, smaller than, equal to
establishes a time series	change, rise, growth, increase, decrease, decline, fluctuation
determines a frequency	range, concentration, most of, distribution of x and y by age
analyses relationships between variables	increase with, decrease with, vary with, despite, correspond to, relate to

4. **Experiment with different types of graphs and select the most appropriate**

- Pie chart (description of components)
- Bar chart (comparison of items and relationships, time series, frequency distribution)
- Line chart (time series and frequency distribution)
- Scatterplot (analysis of relationships)

5.2 Bar chart

A bar chart may be either horizontal or vertical. The important point to note about bar charts is their bar length or height—the greater their length or height, the greater their value. Bar charts are one of the many techniques used to present data in a visual form so that the reader may readily recognize patterns or trends.

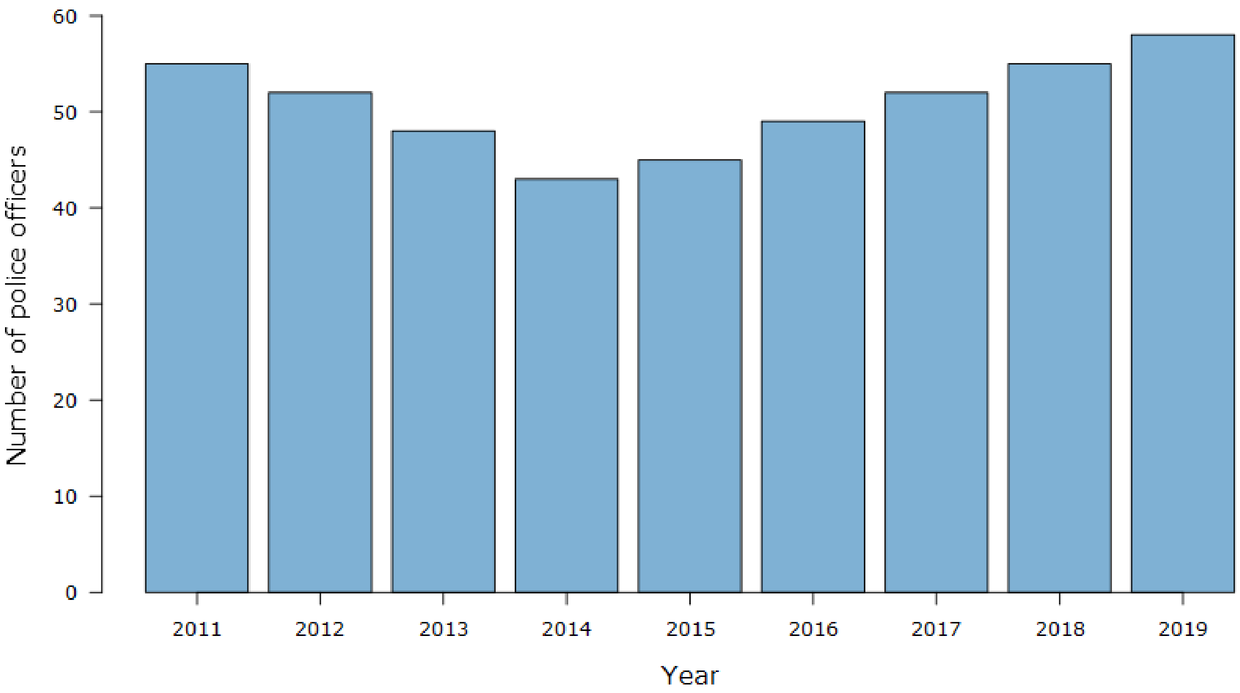
Bar charts usually present categorical variables, discrete variables or continuous variables grouped in class intervals. They consist of an axis and a series of labelled horizontal or vertical bars. The bars depict frequencies of different values of a variable or simply the different values themselves. The numbers on the y-axis of a vertical bar chart or the x-axis of a horizontal bar chart are called the scale.

When developing bar charts manually, draw a vertical or horizontal bar for each category or value. The height or length of the bar will represent the number of units or observations in that category (frequency) or simply the value of the variable. Select an arbitrary but consistent width for each bar as well. Even though it is very common today that some software, such as a spreadsheet software or R software, is used to produce charts, it's still quite useful to know how to create charts by hand.

Vertical bar charts

Bar charts should be used when you are showing segments of information. Vertical bar charts are useful to compare different categorical or discrete variables, such as age groups, classes, schools, etc., as long as there are not too many categories to compare. They are also very useful for time series data. The space for labels on the x-axis is small, but ideal for years, minutes, hours or months. For example, Chart 5.2.1 below shows the number of police officers in Crimeville for each year from 2011 to 2019.

Chart 5.2.1
Number of police officers in Crimeville, 2011 to 2019



In Chart 5.2.1, you can see that the number of police officers decreased from 2011 to 2014, but started increasing again in 2015. The chart also makes it easy to compare the number of police officers for any combination of years.

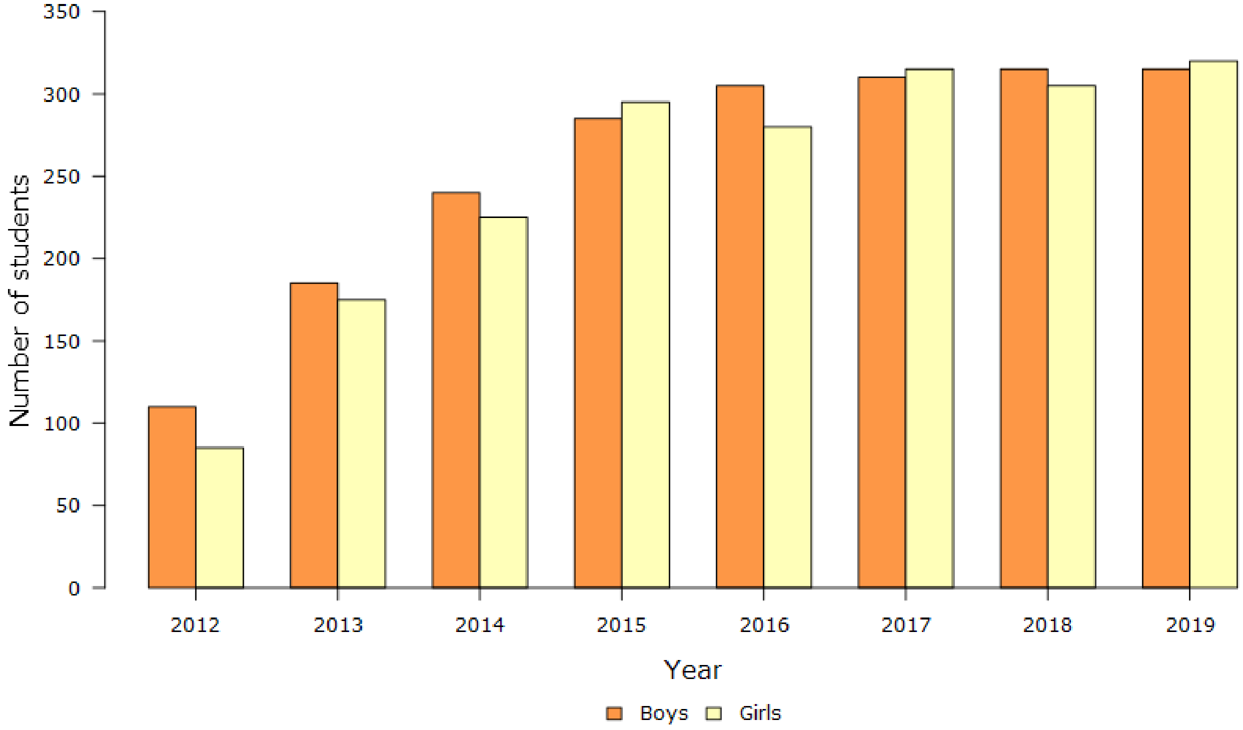
Vertical bar charts are an excellent choice to emphasize a change in magnitude. The best information for a vertical bar chart is data dealing with the description of components, frequency distribution and time-series statistics.

Grouped bar charts

The grouped bar chart is another effective means of comparing sets of data about the same places or items. It gives two or more pieces of information for each item on the x-axis instead of just one as in Chart 5.2.1. This allows you to make direct comparisons on the same chart by age group, gender or anything else you wish to compare. However, if a grouped bar chart has too many series of data, the chart becomes cluttered and it can be confusing to read.

Chart 5.2.2, a grouped vertical bar chart, compares two series of data: the numbers of boys and girls that have a smartphone at Redwood Secondary School from 2012 to 2019. The orange bar represents the number of boys, and the yellow bar represents the number of girls.

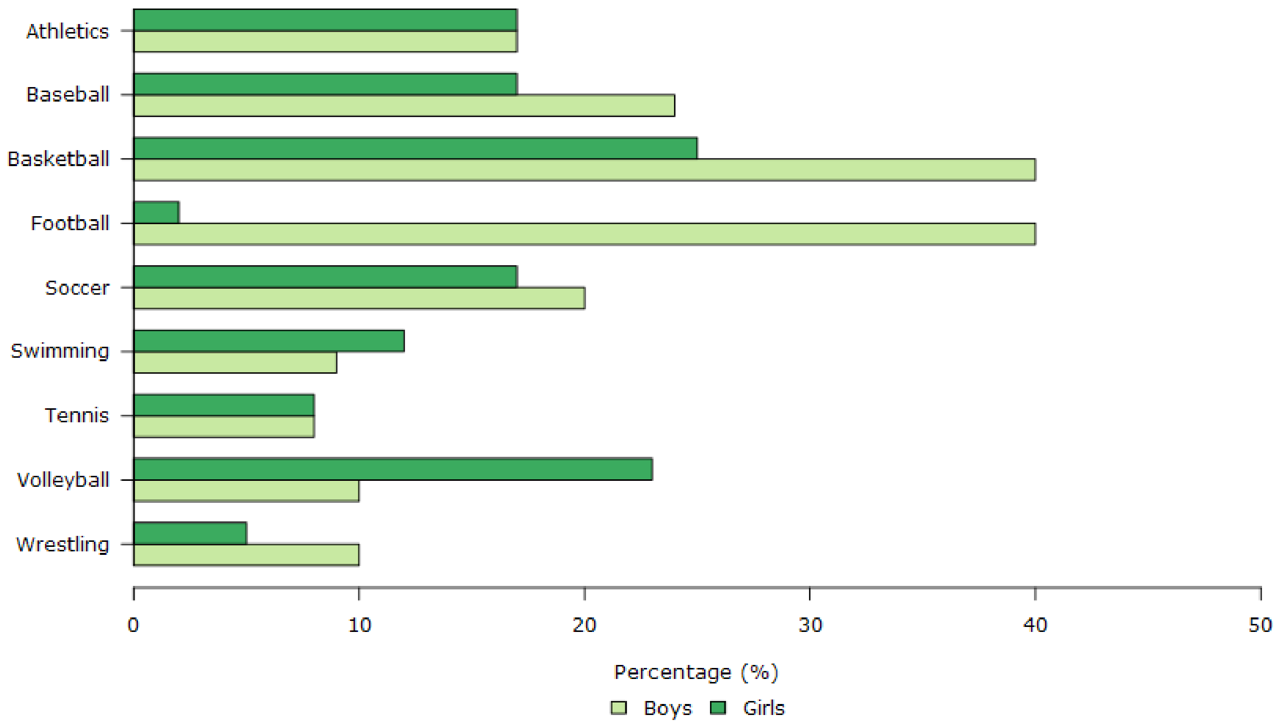
Chart 5.2.2
Students who own a smartphone at Redwood School, by gender, 2012 to 2019



Horizontal bar charts

One disadvantage of vertical bar charts, however, is that they lack space for text labelling at the foot of each bar. When category labels in the chart are too long, you might find a horizontal bar chart better for displaying information, like the example in Chart 5.2.3.

Chart 5.2.3
Sports practiced by 15-year-old students in Jamie's school, by gender



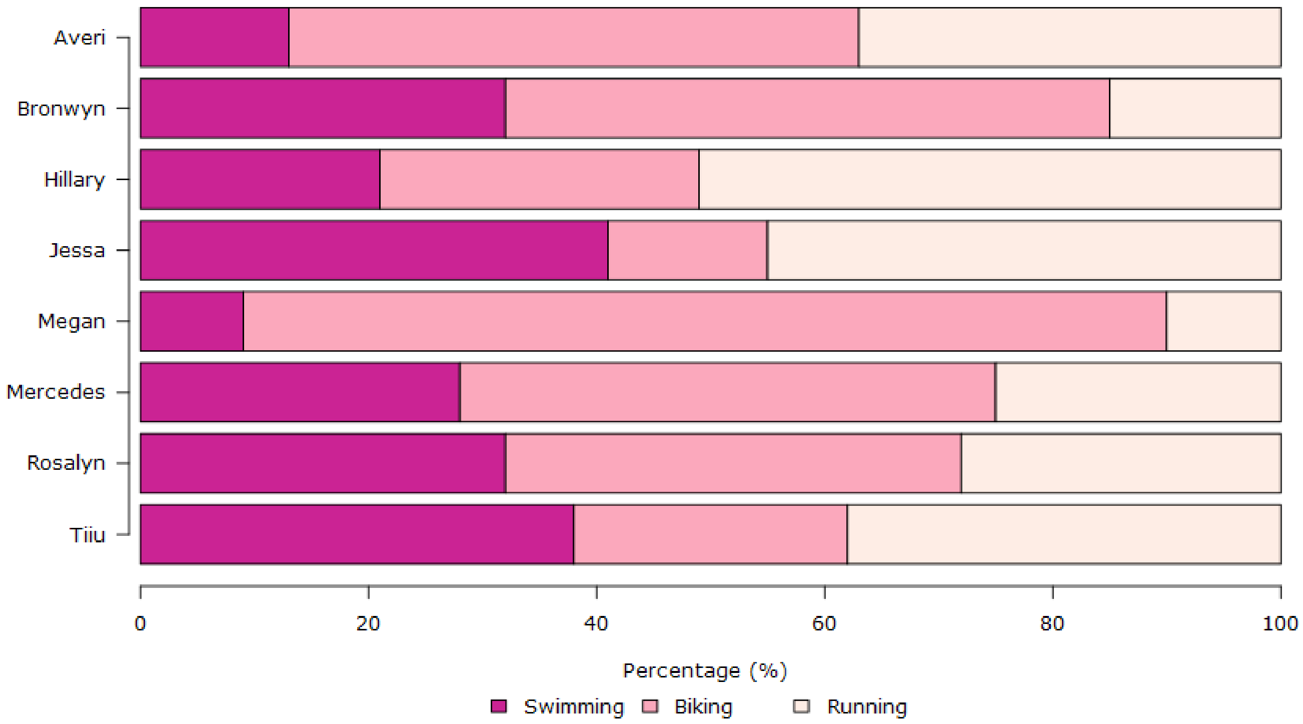
Stacked bar charts

There are several other types of bar chart that you may encounter. The population pyramid is a special application of a grouped bar chart. Another useful type of bar chart is the stacked bar chart.

The stacked bar chart is a preliminary data analysis tool used to show segments of totals. The stacked bar chart can be very difficult to analyze if too many items are in each stack. It can contrast values, but not necessarily in the simplest manner.

In Chart 5.2.4, it is easy to analyze the data presented since there are only three items in each stack: swimming, running and biking. It is easy to see at a glance what percentage of time each woman spent on an event. Had this been a chart representing a decathlon (with 10 events) the data would have been significantly harder to analyze.

Chart 5.2.4
Campbell High Triathlon, percentage of time spent on each event, by competitor



Advices to build bar charts

You should keep the following guidelines in mind when creating bar charts:

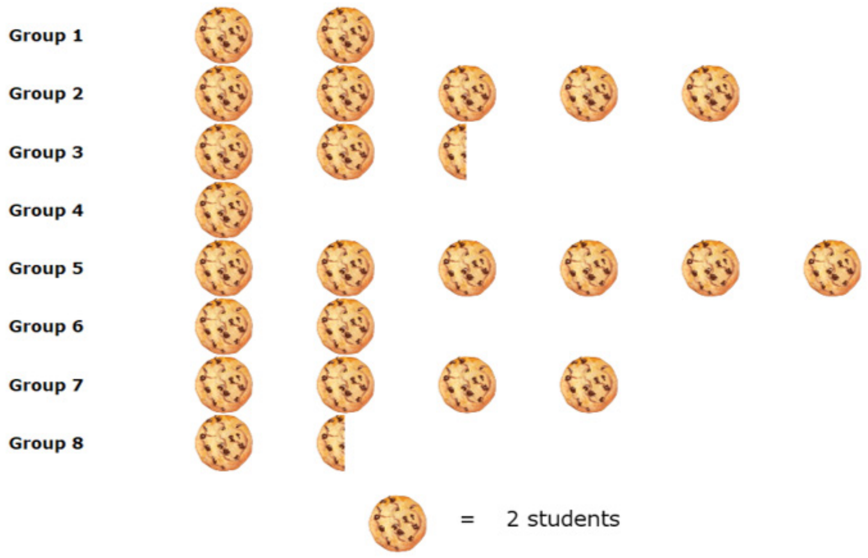
- Make bars and columns wider than the space between them.
- Use a single font type on a chart. Try to maintain a consistent font style from chart to chart in a single presentation or document.
- Order your shade pattern from darkest to lightest.
- Avoid using a combination of red and green in the same display.

5.3 Pictograph

A pictograph uses picture symbols to illustrate statistical information. It is often more difficult to visualize data precisely with a pictograph. This is why pictographs should be used carefully to avoid misrepresenting data either accidentally or deliberately.

Chart 5.3.1 shows a scale that represents the number of elementary students who prefer chocolate chip cookies. This type of pictograph shows how a symbol can be used to represent data. One cookie symbol represents two students, and a half-cookie symbol is used to represent one student. These data could easily have been presented in a bar chart using a scale to present the figure rather than a symbol.

Chart 5.3.1
Number of students who like chocolate chip cookies best



Now let us look at another example of a pictograph.

Chart 5.3.2
Purchasing power of the Canadian dollar, 2000 to 2020

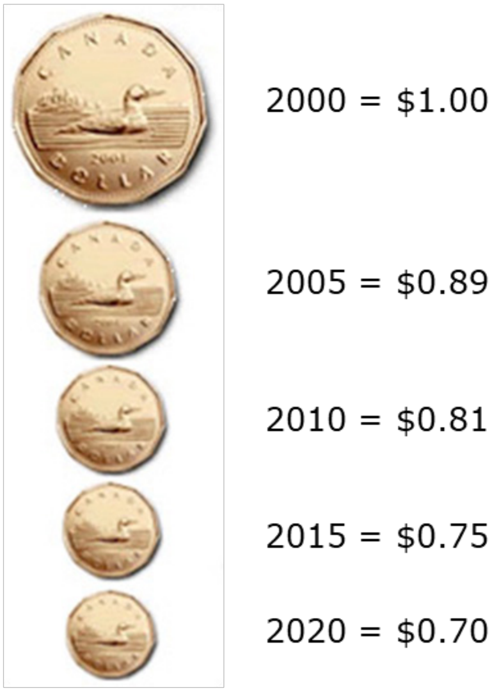


Chart 5.3.2 shows how the Canadian dollar shrank to a value of 70 cents over 20 years because of inflation. This information means the value of the 2020 Canadian dollar was worth 70% of the value of the 2000 Canadian dollar! What is the problem with the depiction of statistics in this pictograph?

The size or area (total surface) of the dollars coin (loonie) pictograph is misleading. The dollar value differences represented are exaggerated by the pictures. They should reflect the actual purchasing power of the dollar of the years in question. Since 70 cents is over half of one dollar, the 2020 loonie should appear bigger than half the size of the 2000 loonie, which is not the case here.

You may argue that people do not notice this misrepresentation when they look at a pictograph such as this one, and thus it is not particularly important. The fact is that subconsciously many people may interpret the Canadian dollar to have lost far more of its value than it has in reality. Since many people use statistical information in making decisions, accuracy is important. In this case, the shrinking value of the Canadian dollar can affect people's perception about their ability to save money or their confidence in the Canada's economy.

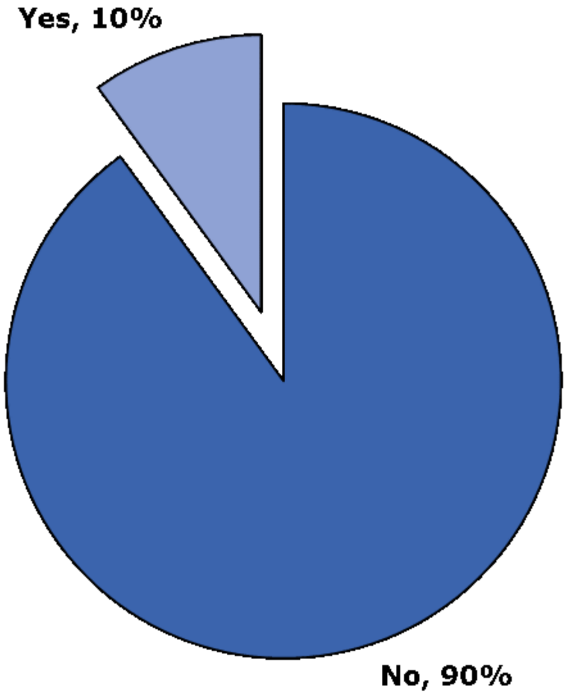
If not drawn carefully, pictographs can be inaccurate. Statistics Canada rarely uses pictographs to release statistical information, but the media uses them quite frequently.

5.4 Pie chart

A pie chart, sometimes called a circle chart, is a way of summarizing a set of nominal data or displaying the different values of a given variable (e.g. percentage distribution). This type of chart is a circle divided into a series of segments. Each segment represents a particular category. The area of each segment is the same proportion of a circle as the category is of the total data set.

Pie chart usually shows the component parts of a whole. Sometimes you will see a segment of the drawing separated from the rest of the pie in order to emphasize an important piece of information. This is called an exploded pie chart. Chart 5.4.1 is an example of an exploded pie chart.

Chart 5.4.1
Student and faculty response to the poll "Should Avenue High School adopt student uniform?"



The pie chart above clearly shows that 90% of all students and faculty members at Avenue High School do not want to have a uniform dress code and that only 10% of the school population would like to have one. This point is clearly emphasized by its visual separation from the rest of the pie.

The use of the pie charts is quite popular, as the circle provides a visual concept of the whole (100%). Pie charts are also one of the most commonly used charts because they are simple to use. Despite its popularity, pie charts should be used sparingly for two reasons. First, they are best used for displaying statistical information when there are no more than six components only—otherwise, the resulting picture will be too complex to understand. Second, pie charts are not useful when the values of each component are too similar because it is difficult to see the differences between slice sizes.

A pie chart uses percentages to compare information. Percentages are used because they are the easiest way to represent a whole. The whole is equal to 100%. For example, if you spend 7 hours at school and 55 minutes of that time is spent eating lunch, then 13.1% of your school day was spent eating lunch. To present this in a pie chart, you would need to find out how many degrees represent 13.1%. This calculation is done by developing the equation:

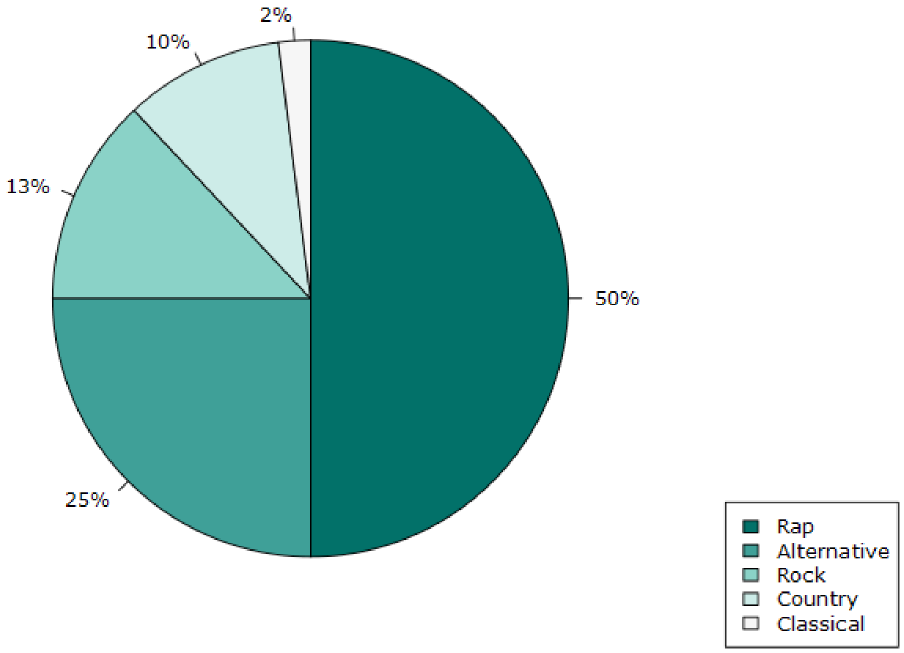
$$\text{percent} \div 100 \times 360 \text{ degrees} = \text{the number of degrees}$$

This ratio works because the total percent of the pie chart represents 100% and there are 360 degrees in a circle. Therefore 47.1 degrees of the circle (13.1%) represents the time spent eating lunch.

Constructing a pie chart

A pie chart is constructed by converting the share of each component into a percentage of 360 degrees. In Chart 5.4.2, music preferences in 14- to 19-year-olds are clearly shown.

Chart 5.4.2
Music genres preferred by young adults 14 to 19



The pie chart quickly tells you that

- half of students like rap best (50%), and
- the remaining students prefer alternative (25%), rock and roll (13%), country (10%) and classical (2%).

Tip! When drawing a pie chart, ensure that the segments are ordered by size (largest to smallest) and in a clockwise direction.

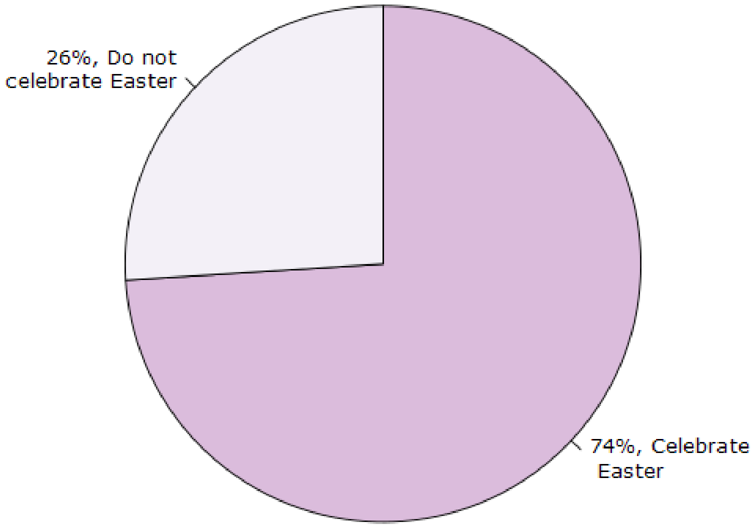
In order to reproduce this pie chart, follow this step-by-step approach:

If 50% of the students liked rap, then 50% of the whole circle graph (360 degrees) would equal 180 degrees.

- Draw a circle with your protractor.
- Starting from the 12 o'clock position on the circle, measure an angle of 180 degrees with your protractor. The rap component should make up half of your circle. Mark this radius off with your ruler.
- Repeat the process for each remaining music category, drawing in the radius according to its percentage of 360 degrees. The final category need not be measured as its radius is already in position.

Labelling the segments with percentage values often makes it easier to tell quickly which segment is bigger. If there are few categories, the percentage and the category label should be indicated beside their corresponding segments like in Chart 5.4.3. This way, users do not have to constantly look back at the legend in order to identify what category each colour represents.

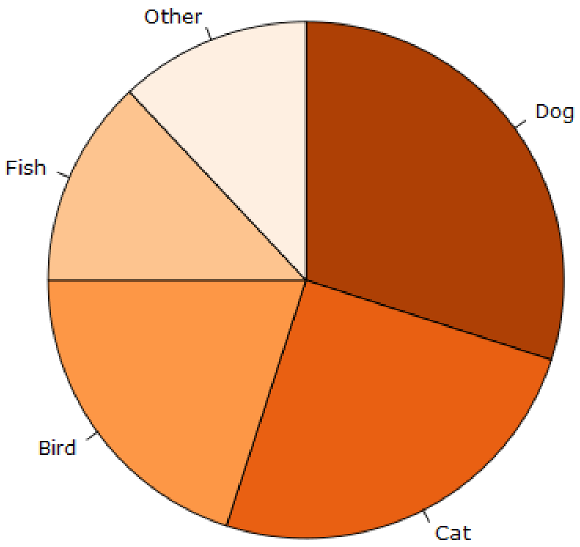
Chart 5.4.3
Percentage of students in Mr. Paul's World Religions class who celebrate Easter



The pie chart above conveys a clear message to the user—that 74% of all students in the World Religions class celebrate Easter. We can easily tell what the message is by simply looking at the accompanying percentages.

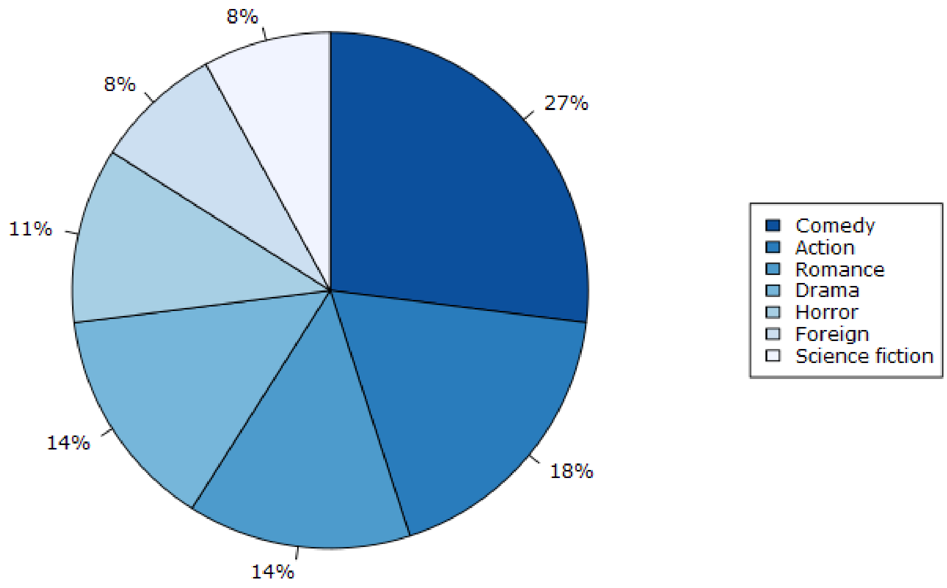
It is more difficult to understand the message behind Chart 5.4.4 because there are no percentage figures given for each slice of the pie. The user can still develop a picture of what is being said about the type of pets sold by this store, but the message is not as clear as it would have been had the parts of the pie been labelled.

Chart 5.4.4
Pets bought at World of Pets



In the Chart 5.4.5 below, the legend is formatted properly and the percentages are included for each of the pie segments. However, there are too many items in the pie chart to quickly give a clear picture of the distribution of movie genres. If there are more than five or six categories, consider using another type of graph to display the information. Chart 5.4.5 would certainly be easier to read as a bar chart.

Chart 5.4.5
Favourite movie genres in Mrs Smyth's Film class



> **Tip!** Many software, like spreadsheets, will draw pie charts for you quickly and easily. However, research has shown that many people can make mistakes when trying to compare pie chart values. In general, bar charts communicate the same message with less chance for misunderstanding.

Pie charts versus bar charts

When displaying statistical information, refrain from using more than one pie chart for each figure. Chart 5.4.6 shows two pie charts side-by-side, where a grouped bar chart would have shown the information more clearly. A user might find it difficult to compare a segment from one pie chart to the corresponding segment of the other pie chart. However, in a grouped bar chart, these segments become bars which are lined up side by side, making it much easier to make comparisons.

Chart 5.4.6
Smoking frequency of 15-year-olds on the Parkview Secondary School track and field team, by gender

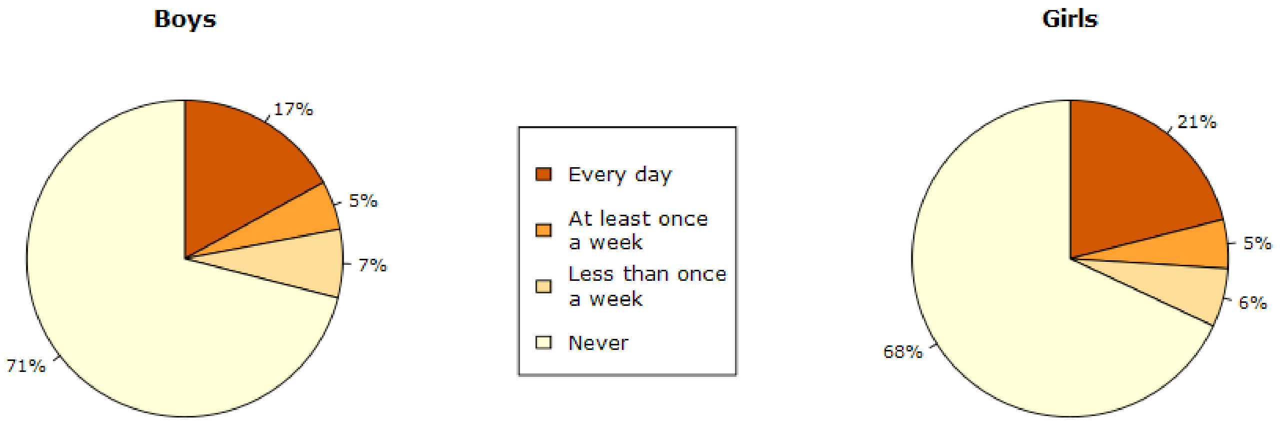
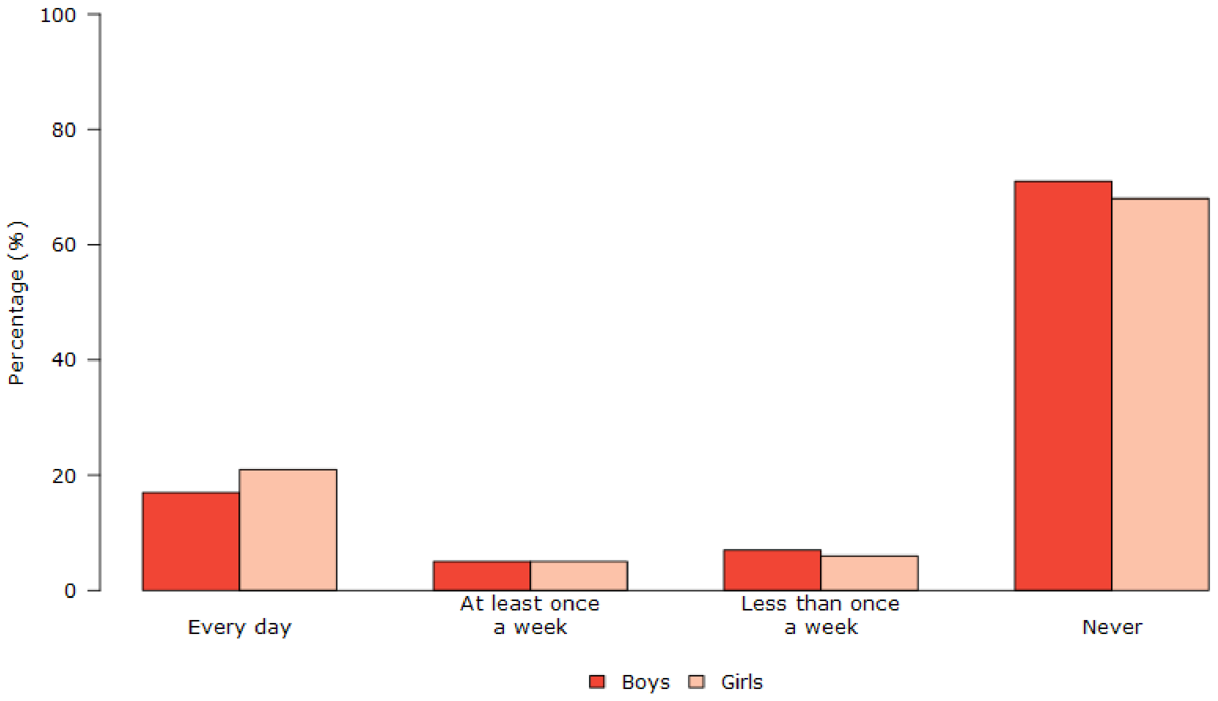


Chart 5.4.7 shows how a grouped bar chart would be a better choice for displaying information than a double pie chart. The key point in preparing this type of graph is to ensure that you are using the same scale for both categories of the bar chart. You'll notice that the information is much clearer in Chart 5.4.7 than in Chart 5.4.6.

Chart 5.4.7
Smoking frequency of 15-year-olds on the Parkview Secondary School track and field team, by gender



5.5 Line chart

Line charts, especially useful in the fields of statistics and science, are more popular than all other graphs combined because their visual characteristics reveal data trends clearly and these charts are easy to create.

A line chart is a visual comparison of how two variables—shown on the x- and y-axes—are related or vary with each other. It shows related information by drawing a continuous line between all the points on a grid.

Line charts compare two variables: one is plotted along the x-axis (horizontal) and the other along the y-axis (vertical). The y-axis in a line chart usually indicates quantity (e.g. dollars, litres) or percentage, while the horizontal x-axis often measures units of time. As a result, the line chart is often viewed as a time series graph. For example, if you wanted to graph the height of a baseball pitch over time, you could measure the time variable along the x-axis, and the height along the y-axis. Although they do not present specific data as well as tables do, line charts are able to show relationships more clearly than tables do. Line charts can also depict multiple series and hence are usually the best candidate for time series data and frequency distribution.

Vertical bar charts and line charts share a similar purpose. The vertical bar chart, however, reveals a change in magnitude, whereas the line chart is used to show a change in direction.

In summary, line charts:

- show specific values of data well,
- reveal trends and relationships between data,
- compare trends in different groups.

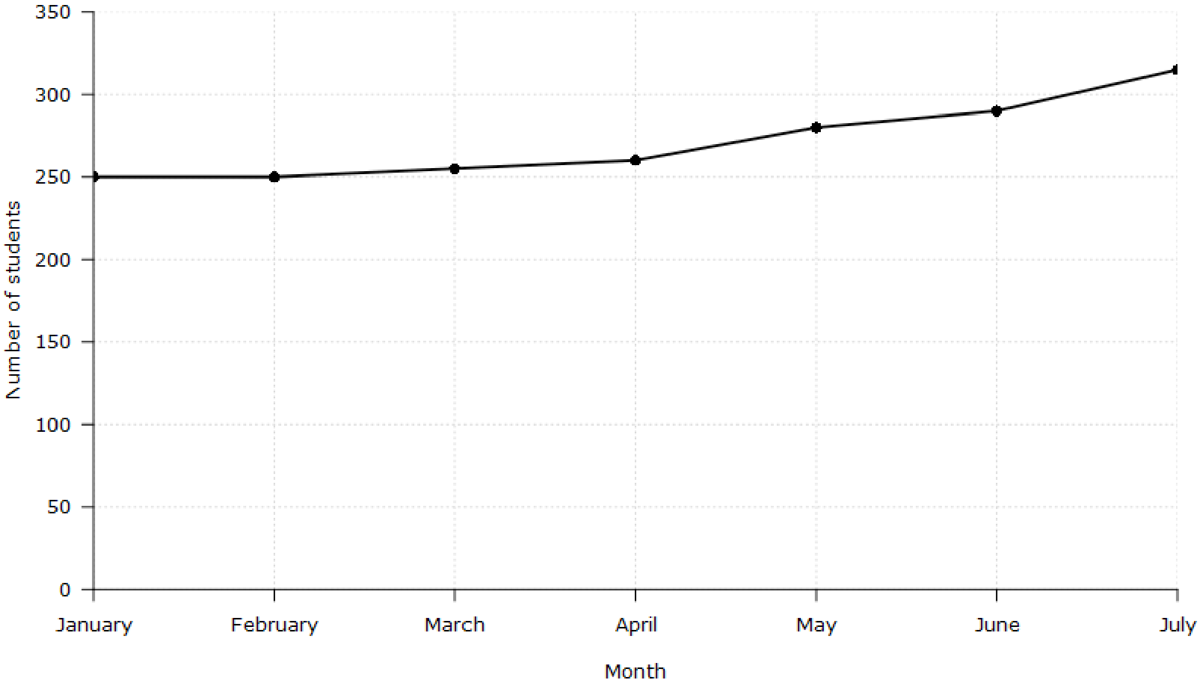
Graphs can give a distorted image of the data. If scales on the axes of a line graph force data to appear a certain way, then a graph can even reveal a trend that is entirely different from the one intended. This happens when the intervals between adjacent points along the axis may be dissimilar, or when the same data charted in two graphs using different scales appear different.

Example 1 – Plotting a trend over time

Chart 5.5.1 shows one obvious trend, the fluctuation in the labour force from January to July. The number of students at Andrew's high school who are members of the labour force is scaled using intervals on the y-axis, while the time variable is plotted on the x-axis.

The number of students participating in the labour force was 252 in January, 252 in February, 255 in March, 256 in April, 282 in May, 290 in June and 319 in July. When examined further, the line chart indicates that the labour force participation of these students was at a plateau for the first four months (January to April), and for the next three months (May to July) the number increased steadily.

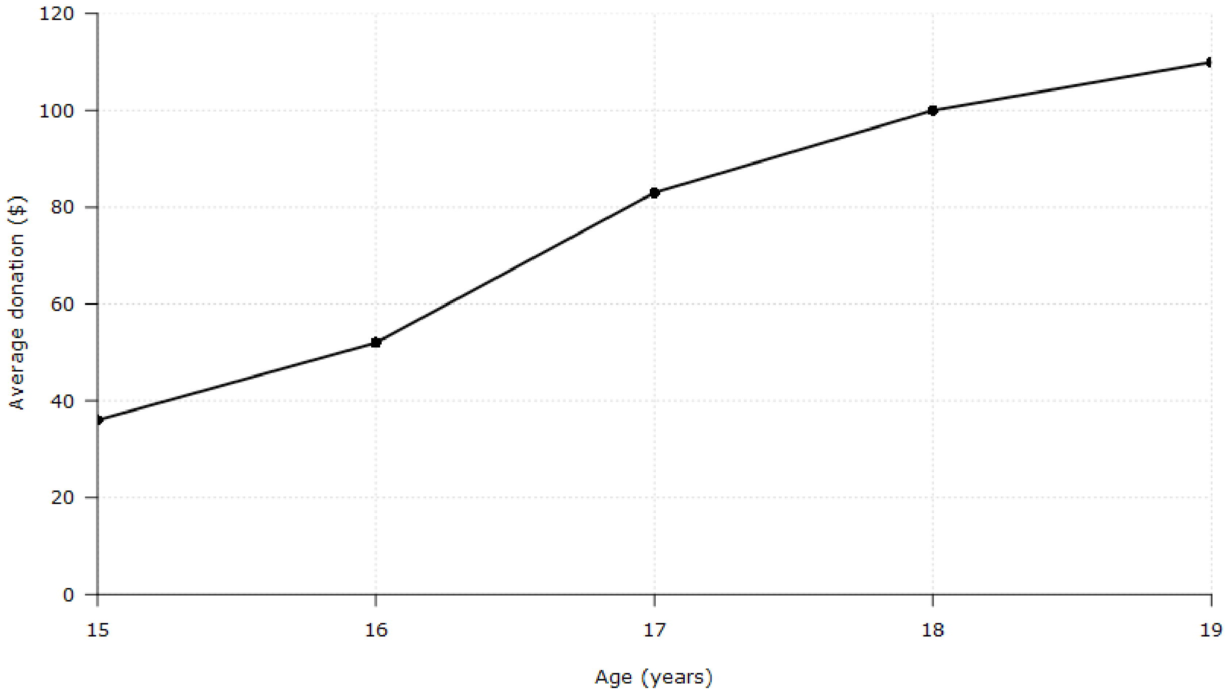
Chart 5.5.1
Labour force participation in Andrew's high school



Example 2 – Comparing two related variables

Chart 5.5.2 is a single line chart comparing two items. In this example, time is not a factor. The chart compares the average number of dollars donated by the age of the donors. According to the trend in the chart, the older the donor, the more money he or she donates. The 17-year-old donors donate, on average, $84. For the 19-year-olds, the average donation increased by $26 to make the average donation of that age group $110.

Chart 5.5.2
Average number of dollars donated at Evergreen High School, by age of the donors



Example 3 – Using correct scal

When drawing an axis, it is important that you use the correct scale. Otherwise, the line's shape can give readers the wrong impression about the data. Compare Chart 5.5.3 with Chart 5.5.4:

Chart 5.5.3
Number of guilty crime offenders, Grishamville

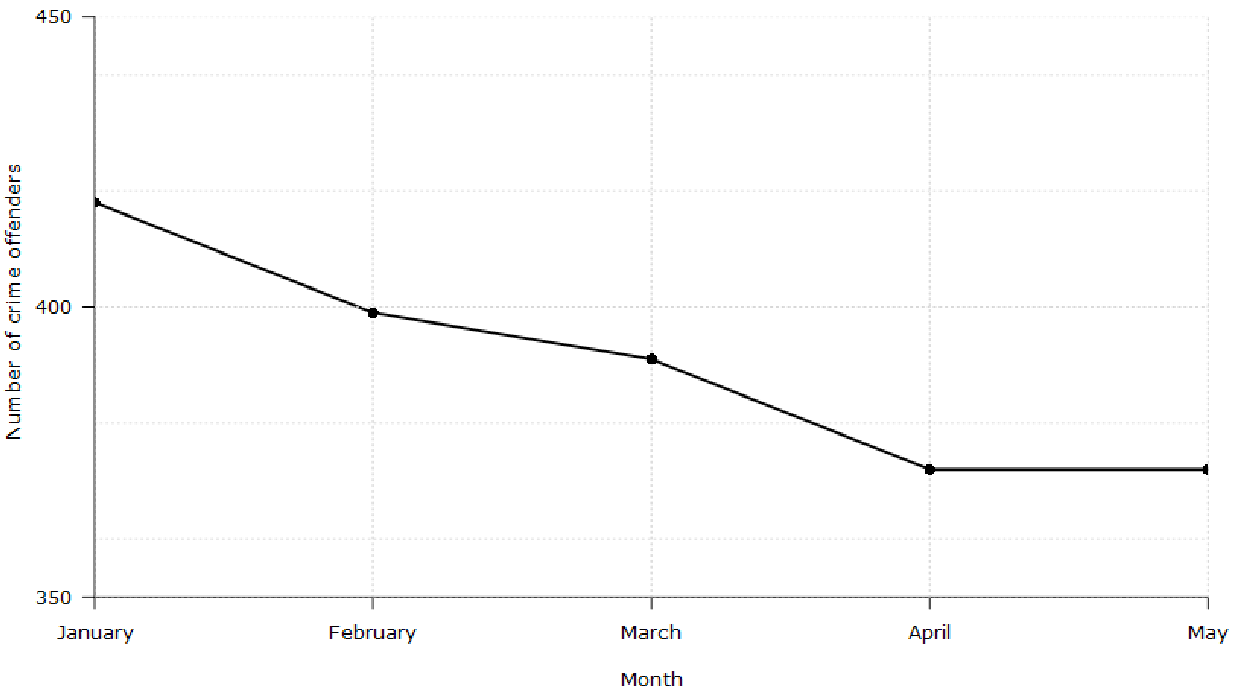
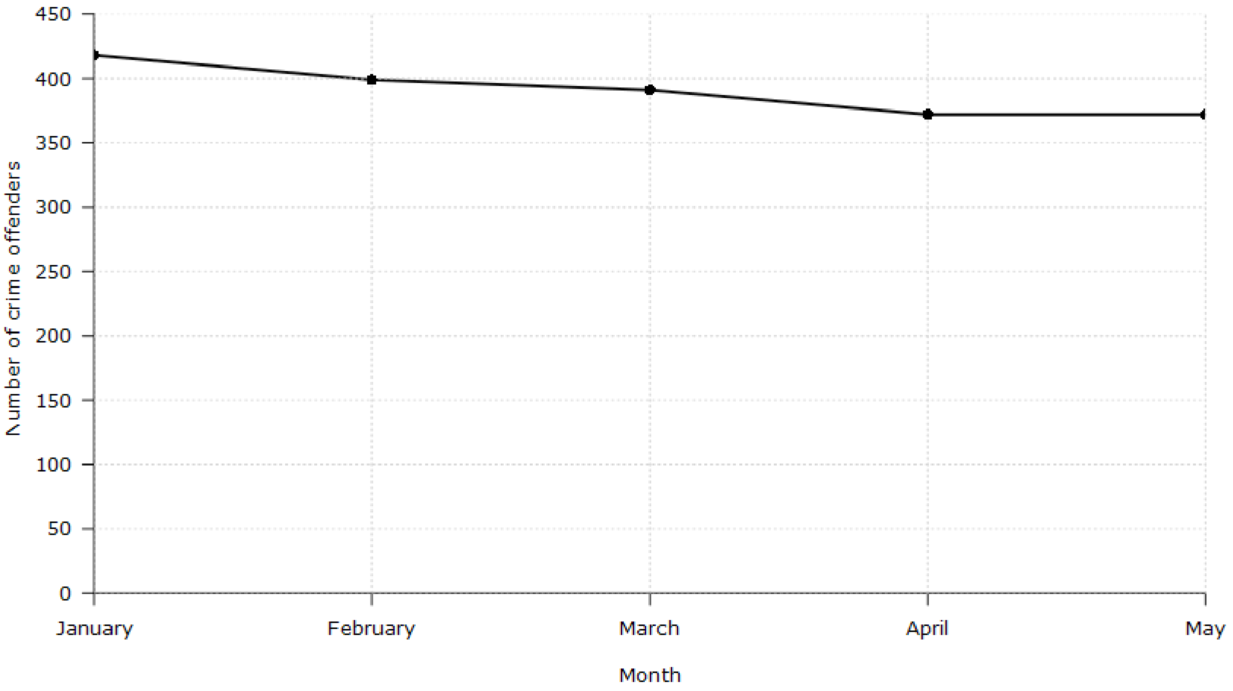


Chart 5.5.4
Number of guilty crime offenders, Grishamville



Using a scale of 350 to 430 (Chart 5.5.3) focuses on a small range of values. It does not accurately depict the trend in guilty crime offenders between January and May since it exaggerates that trend. However, choosing a scale of 0 to 450 (Chart 5.5.4) better displays how small the decline in the number of guilty crime offenders really was.

Both charts can be useful depending on the context. The important thing to remember is that you should pay attention to the scale that is being used when interpreting a graph.

Example 4 – Multiple line graphs

A multiple line chart can effectively compare similar items over the same period of time, as you can see in Chart 5.5.5 which compares the use of cell phones by gender.

Chart 5.5.5
Cell phone use in Anytown by gender, 2012 to 2018

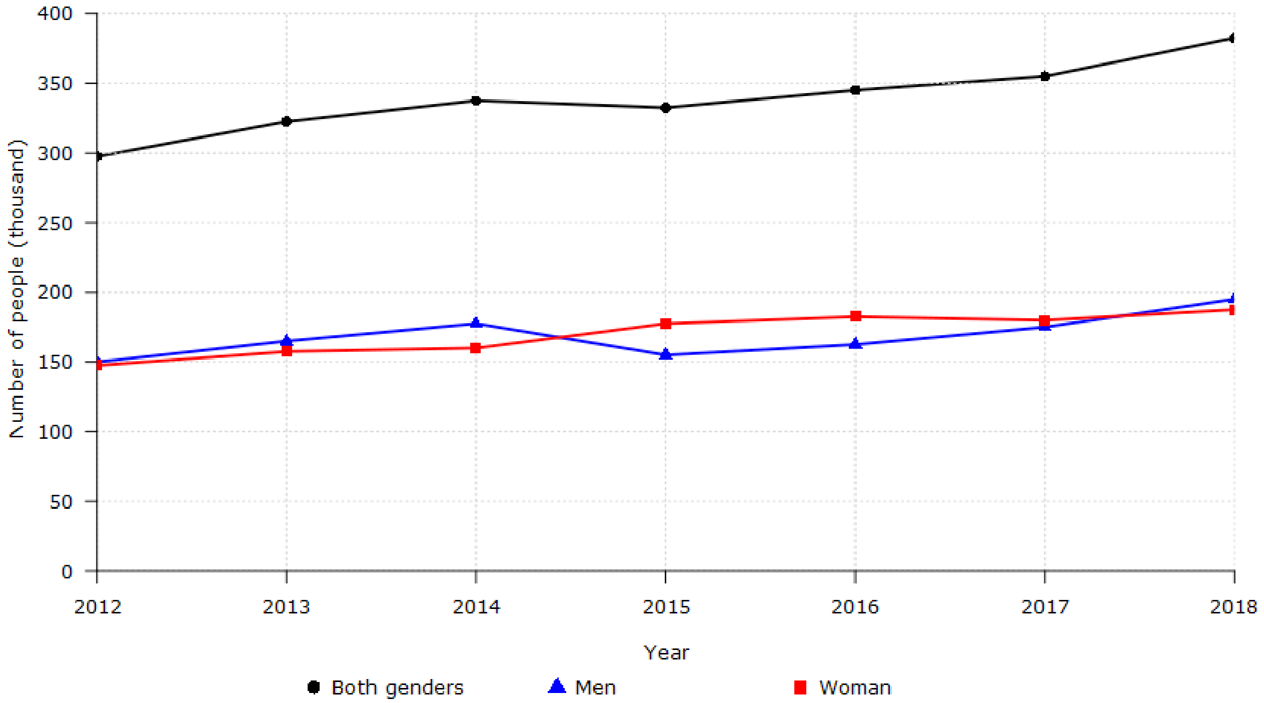


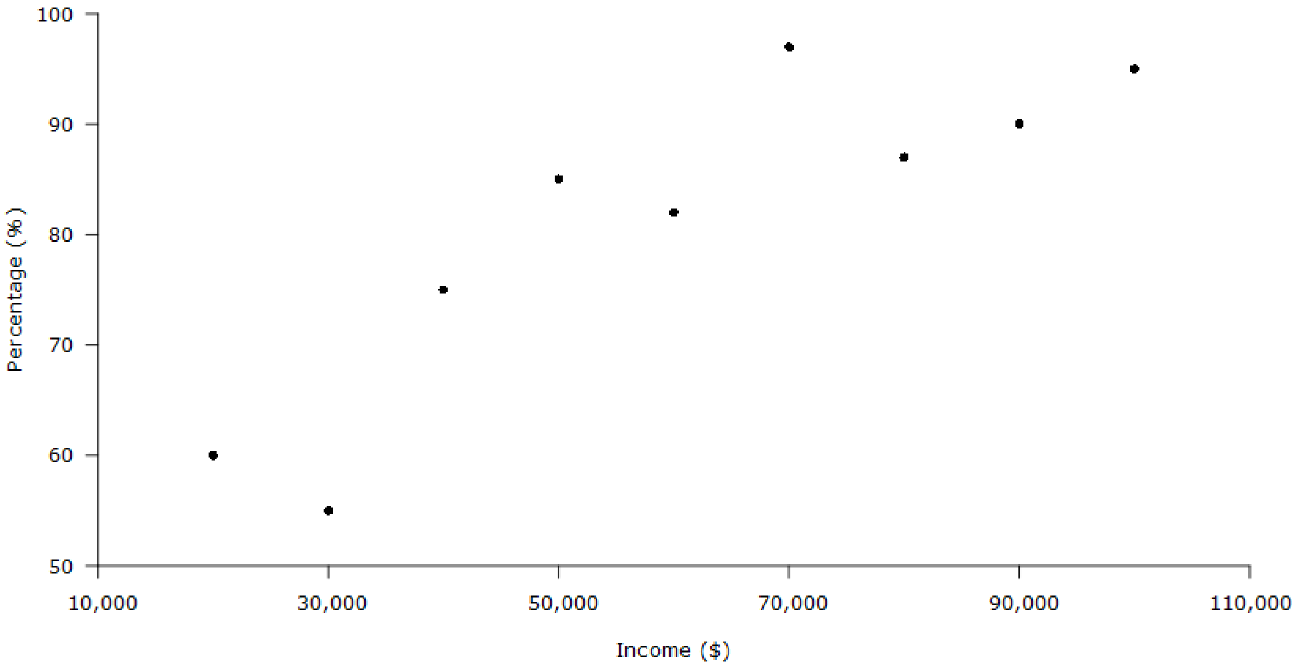
Chart 5.5.5 is an example of a good chart. The message is clearly stated in the title, and each of the line graphs is properly labelled. It is easy to see from this chart that the total cell phone use has been rising steadily since 2012, except for a one-year period (2015) where the numbers drop slightly. The pattern of use for women and men seems to be quite similar with very small discrepancies between them.

5.6 Scatter plot

In science, the scatterplot is widely used to present measurements of two or more related variables. It is particularly useful when the values of the variables of the y-axis are thought to be dependent upon the values of the variable of the x-axis.

In a scatterplot, the data points are plotted but not joined. The resulting pattern indicates the type and strength of the relationship between two or more variables. Chart 5.6.1 is an example of a scatterplot. Car ownership increases as the household income increases, showing that there is a positive relationship between these two variables.

Chart 5.6.1
Car ownership in Anytown, by household income



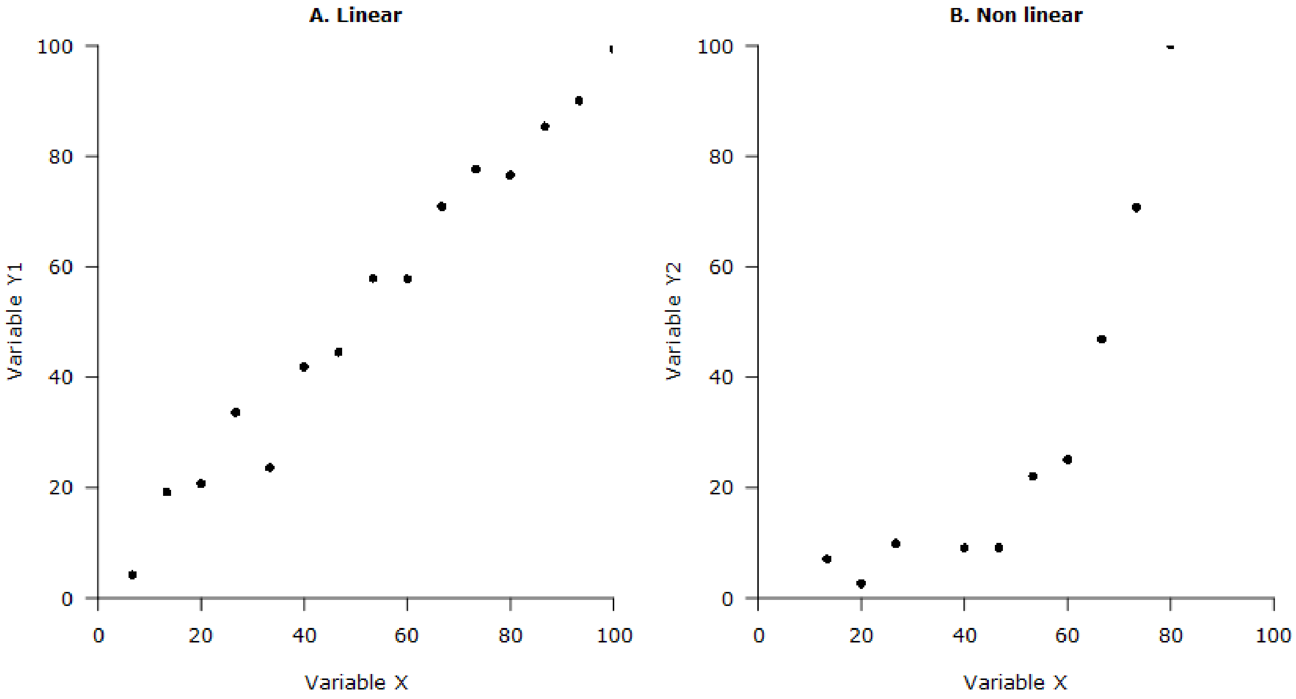
The pattern of the data points on the scatterplot reveals the relationship between the variables. Scatterplots can illustrate various patterns and relationships, such as:

- a linear or non-linear relationship,
- a positive (direct) or negative (inverse) relationship,
- the concentration or spread of data points,
- the presence of outliers.

Linear or non-linear relationship

When the data points form a straight line on the graph, the relationship between the variables is linear, as shown in Chart 5.6.2, Part A. When the data points don't form a line or when they form a line that is not straight, like in Chart 5.6.2, Part B, the relationships between variables is not linear.

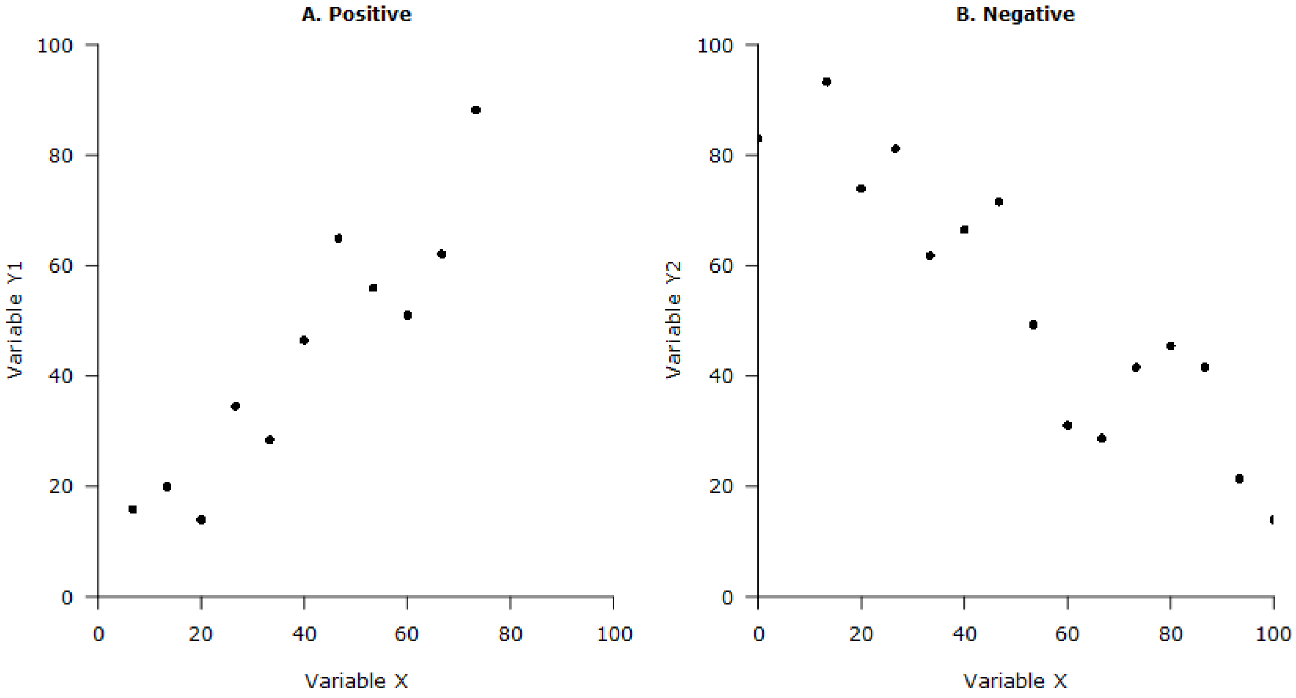
Chart 5.6.2
Linear relation or non linear relationship



Positive or negative relationship

If the points cluster around a line that runs from the lower left to upper right of the graph area, then the relationship between the two variables is said to be positive or direct (Chart 5.6.3, Part A). If the points cluster around a line that runs from the upper left to the lower right of the graph area, then the relationship is said to be negative or inverse (Chart 5.6.3, Part B).

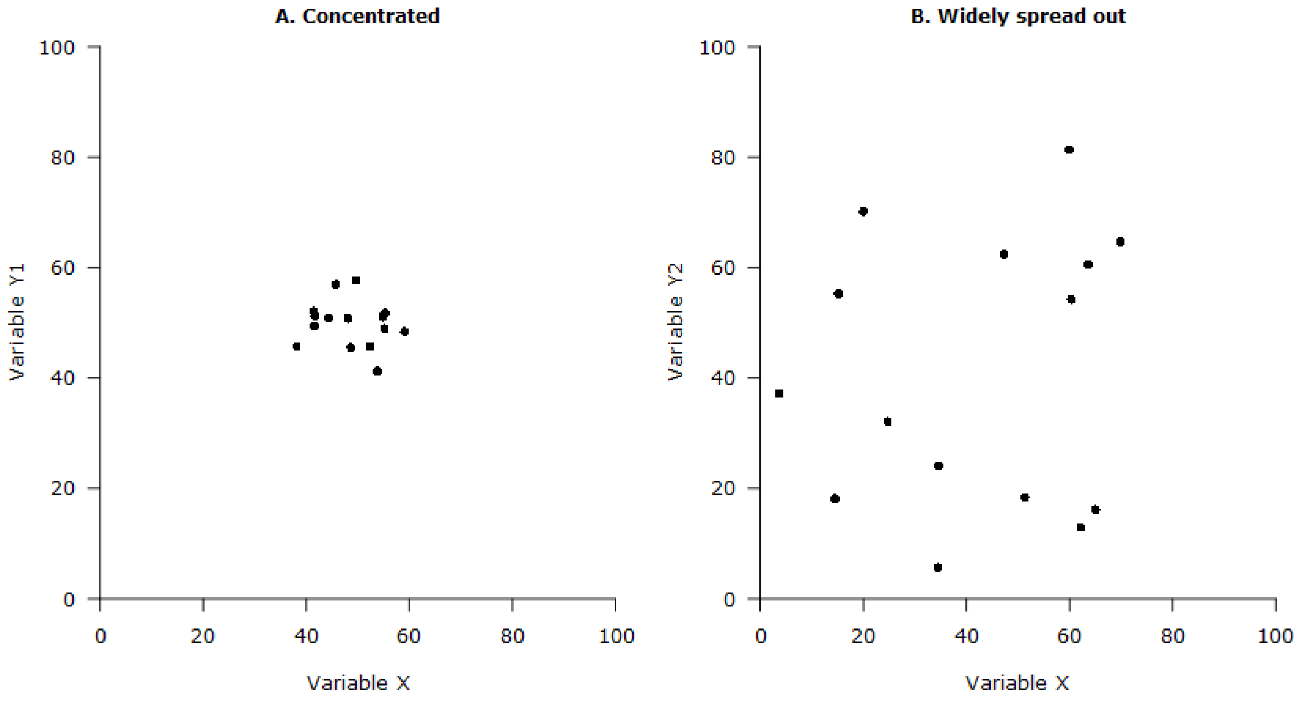
Chart 5.6.3
Positive relation or negative relationship



Concentration or spread of data points

Data points can be close together (Chart 5.6.4, Part A) or spread widely across the graph area (Chart 5.6.4, Part B).

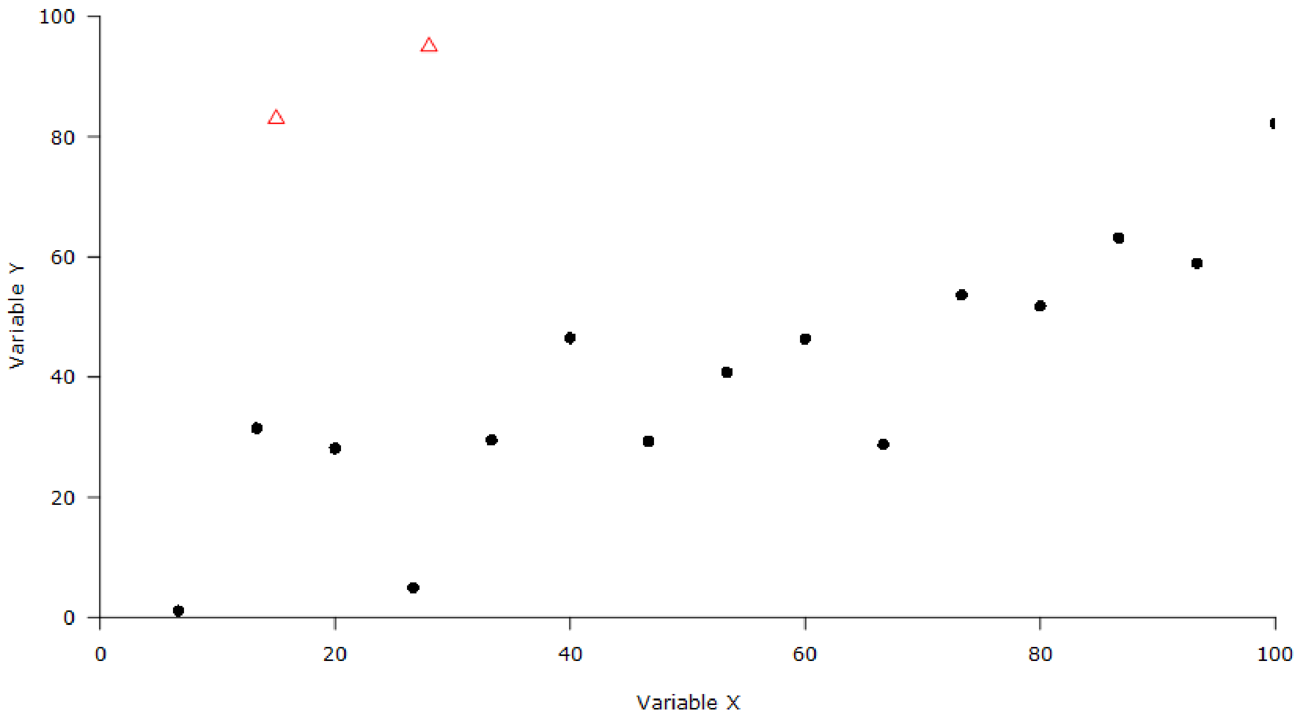
Chart 5.6.4
Concentrated data or widely spread out data



Presence of outliers

Besides portraying relationships between the variables, a scatterplot can also show whether or not there are any outliers in the data. Outliers are data points that are far from the other points in the data set, like the two points in red in Chart 5.6.5.

Chart 5.6.5
Presence of outliers

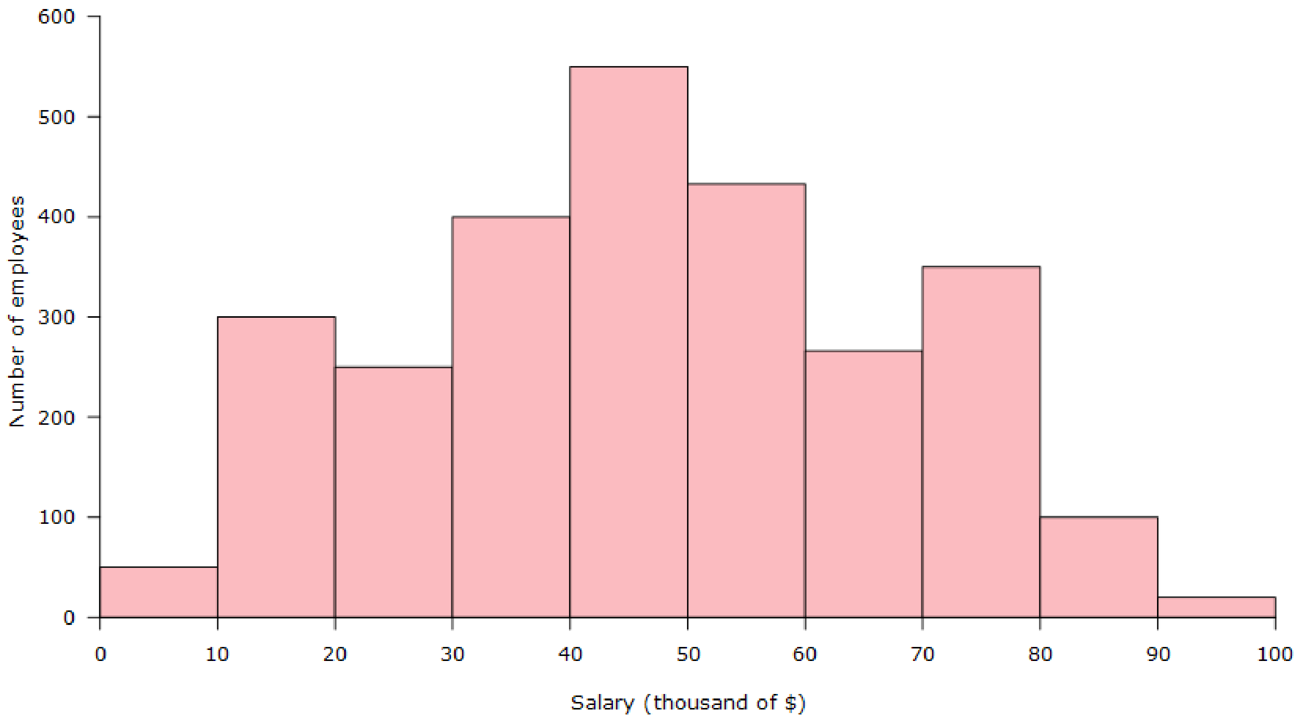


5.7 Histogram

The histogram is a popular graphing tool. It is used to summarize discrete or continuous data that are measured on an interval scale. It is often used to illustrate the major features of the distribution of the data in a convenient form. It is also useful when dealing with large data sets (greater than 100 observations). It can help detect any unusual observations (outliers) or any gaps in the data.

A histogram divides up the range of possible values in a data set into classes or groups. For each group, a rectangle is constructed with a base length equal to the range of values in that specific group and a length equal to the number of observations falling into that group. A histogram has an appearance similar to a vertical bar chart, but there are no gaps between the bars. Generally, a histogram will have bars of equal width. Chart 5.7.1 is an example of a histogram that shows the distribution of salary, a continuous variable, of the employees of a corporation.

Chart 5.7.1
Distribution of salaries of the employees of ABC Corporation



The following table presents the differences between a histogram and vertical bar graph.

Table 5.7.1
Differences between bar chart and histogram

Comparison terms	Bar chart	Histogram
Usage	To compare different categories of data.	To display the distribution of a variable.
Type of variable	Categorical variables	Numeric variables
Rendering	Each data point is rendered as a separate bar.	The data points are grouped and rendered based on the bin value. The entire range of data values is divided into a series of non-overlapping intervals.
Space between bars	Can have space.	No space.
Reordering bars	Can be reordered.	Cannot reordered.

5.8 Exercises

1. The number of basketball games attended by 50 season ticket holders were:
15, 10, 17, 11, 15, 12, 13, 16, 12, 14, 14, 16, 15, 18, 11, 16, 13, 17, 12, 16, 18, 15, 17, 15, 19, 13, 14, 17, 16, 15, 12, 11, 17, 16, 15, 10, 14, 15, 13, 16, 18, 15, 17, 11, 14, 17, 15, 14, 13, 16.

a. Tally the data and present them in a cumulative frequency table.

b. Draw a vertical bar chart.

c. Describe the data set using the five-number summary, the range and the interquartile range. These concepts have been presented in section 4 on data exploration.

5.9 Answers

1. a.

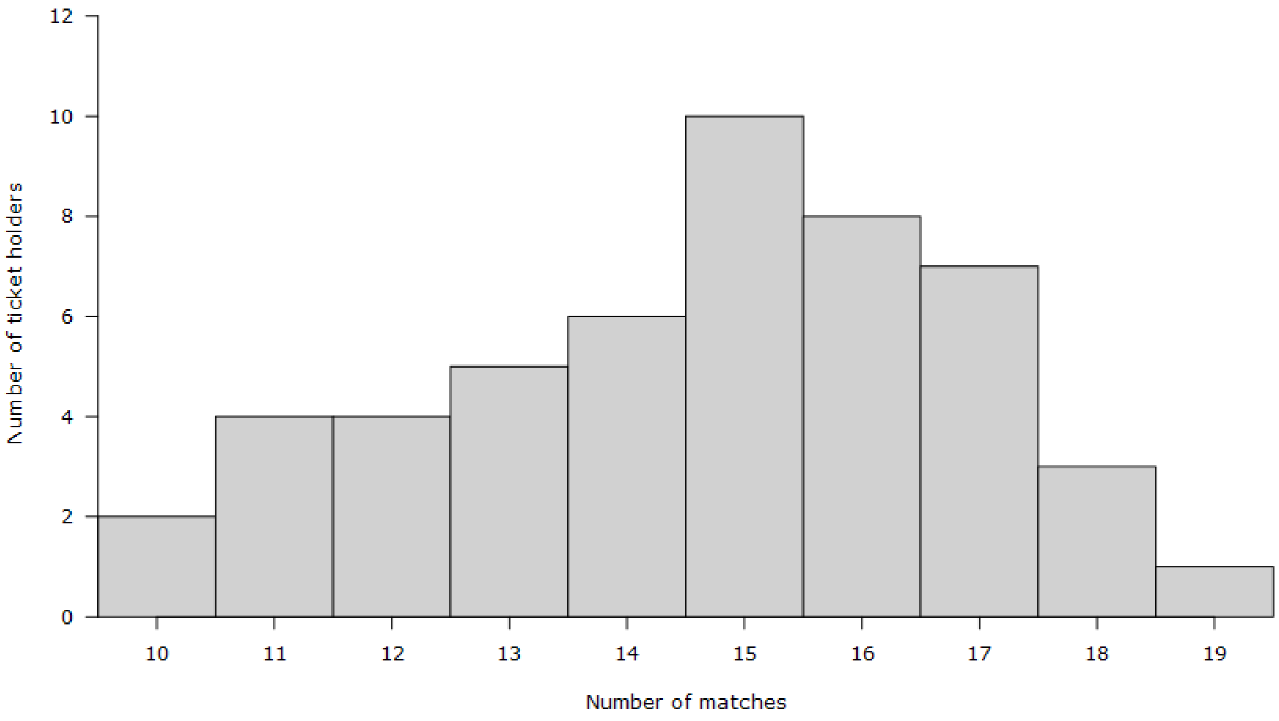
Table 5.9.1
Number of basketball games attended by 50 season ticket holders

Number of matches (x)	Frequency (f)	Percentage (%)	Cumulative Frequency	Cumulative Percentage (%)
10	2	4	2	4
11	4	8	6	12
12	4	8	10	20
13	5	10	15	30
14	6	12	21	42
15	10	20	31	62
16	8	16	39	78
17	7	14	46	92
18	3	6	49	98
19	1	2	50	100
TOTAL	50	100	…	…

... not applicable

b.

Chart 5.9.1
Number of basketball matches attended by 50 season ticket holders



c. The five-number summary is

- Minimum: 10
- Lower quartile: 13
- Median: 15
- Upper quartile: 16
- Maximum: 19

The range is 9 and the interquartile range is 3. More than 50% of the data points are in the interval 13 to 16. The most frequent value (mode) is 15.

Bibliography

Australian Bureau of Statistics (1998). *Statistics – A Powerful Edge!* (2nd ed).

Beaumont, J.-F. (2020). Are probability surveys bound to disappear for the production of official statistics? *Survey Methodology*, Statistics Canada, Catalogue No 12-001-X, Vol. 46, No. 1.

Lorh, S.L. (2019). Sampling: Design and Analysis (2nd ed). Chapman & Hall/CRC Press.

Statistics Canada (2003). *Survey Methods and Practices*. Catalogue no 12-587-XPF.

Statistics Canada (2019). *Frequently asked questions on using new and existing data for official statistics*.

Statistics Canada (2019). *Statistics Canada Quality Guidelines – Sixth edition*. Catalogue no 12-539-X.

Statistics Canada (2020). *Data quality in six dimensions* [Video]. Catalogue no 892000062020001.

Statistics Canada (2020). *Types of Data: Understanding and Exploring Data* [Video]. Catalogue no 892000062020004.

Statistics Canada (2020). *What is Data? An Introduction to Data Terminology and Concepts* [Video]. Catalogue no 892000062020006.

Statistics Canada (2021). *Statistics 101: Correlation and causality* [Video]. Catalogue no 892000062021002.

Statistics Canada (2021). *Statistics 101: Exploring measures of central tendency* [Video]. **Catalogue no** 892000062020002.

Statistics Canada (2021). *Statistics 101: Exploring measures of dispersion* [Video]. Catalogue no 892000062020003.

Statistics Canada (2021). *Statistics 101: Proportion, ratios and rates* [Video]. Catalogue no 892000062021003.

Glossary

The definitions below provide information for those who have questions about some terms used in statistics, but who do not need highly technical definitions . These definitions provided here are, in some cases, oversimplifications of highly complex concepts. For more detailed explanations, you can consult the references provided one the Bibliography page.

A

Administrative data

Data collected as a result of an organization's day-to-day operations.

Aggregate data

Data set in which one record represents a summary of multiple observation units.

B

Big data

Data sets that have such a large number of records and variables that they exceed the capacity of traditional software to process the information within a reasonable time.

Box and whisker plot

Type of graph used to visualize the five-number summary, i.e. the median, the lower and upper quartiles, the minimum and the maximum. **Synonym: box plot**.

C

Categorical variable

Characteristic that isn't quantifiable. **Synonym: qualitative variable**.

Census

In general, survey that aims to collect information about every unit of a population. A census is also used to list and count all units of a population.

Central tendency

Measure of the location of the middle or the centre of a distribution.

Closed question

In a questionnaire, a closed question gives the respondent a list of predefined answers and the respondent is supposed to select one or more answers from the list.

Coefficient of variation

Ratio of the standard error of the estimate to the average value of the estimate across all possible samples.

Confidence interval

The range of values around the estimate that is likely to include the unknown population true value with a given probability.

Continuous variable

Numeric variable that assumes an infinite number of real values within a given interval.

Crowdsourcing

Collection of data information from a large community of users. It relies on the principle that citizens are the experts of their local environment.

D

Data

Facts, figures, observations, or recordings that can take the form of image, sound, text or physical measurements (distance, weight, wave lengths, etc.). Data can be gathered and processed in order to form conclusions.

Data capture

The process used to convert data in a machine-readable format.

Data coding

The process that assigns a value (code) to a response. The code can be a numeric value or a character string.

Data editing

Application of checks to detect missing, invalid or inconsistent values or to point to data records that are potentially in error.

Data imputation

The process used to assign replacement values for missing, invalid or inconsistent data that have failed edits.

Data item

The smallest piece of information that can be gathered from a source of information.

Data processing

Transformation of raw data so they can be used to produce estimates or to carry other data analysis.

Data provider

Individual or organization that collect and process data because information is needed for different purposes, and make these data accessible to data users.

Data set

Grouping of data that have common definitions of observation units and variables.

Database

Structured set of data items, generally presented as tables.

Delimited text file

A text file used to store data, in which each line represents a unit, and each line has fields separated by a delimiter. The most common delimiters are commas, tab, and colon.

Discrete variable

Numeric variable that assumes only a finite number of real values within a given interval. The possible values can be enumerated and counted.

Dispersion

Measure of the spread of a distribution around the central tendency.

F

Frequency

The number of times a value occurs in a data set. It can also be a number of events or items. **Synonym: count**.

Frequency distribution

Chart or table showing how many times each value or range of values of a variable appear in a data set.

I

Interquartile range

Range of the 50% of data that is central to the distribution, i.e. the difference between the upper quartile and the lower quartile.

L

Lower quartile

Value under which 25% of data points are found when they are arranged in increasing order. **Synonym: first quartile**.

M

Margin of error

Half the width of the confidence interval associated to an estimate.

Mean

Measure of central tendency which is the sum of all values divided by the number of values.

Median

Value in the middle of a data set, meaning that 50% of data points have a value smaller or equal to the median and 50% of data points have a value higher or equal to the median. **Synonym: second quartile**.

Metadata

Data about data or data elements, including data descriptions, ownership, access paths, access rights, quality or other information that provides context to data.

Microdata

Data set in which one record represents one unit of observation.

Missing value

Blank or absent data point.

Mode

For categorical or discrete variables, it is the value(s) for which the highest frequency is observed. For continuous variables, the modal-class intervals are the peaks of the histogram. When the mode is unique, it can be used as a measure of central tendency.

N

Nominal variable

Categorical variable that describes a name, label or category without natural order.

Non-sampling errors

All sources of error that are unrelated to sampling.

Numeric variable

A quantifiable characteristic whose values are numbers. **Synonym: quantitative variable**.

O

Open data

Structured, machine-readable data that are freely shared and that can be used without restrictions.

Open question

In a questionnaire, an open question gives the respondent an opportunity to answer the question in their own words.

Ordinal variable

Categorical variable whose values are defined by an order relation between the different categories.

P

Primary source of information

Data from a primary source was collected for the purpose of producing statistics and statistical information.

Q

Questionnaire

Series of questions designed to elicit information on one or more topics from a respondent.

R

Range

Difference between the largest value (maximum) and the smallest value (minimum).

Record linkage

The process by which records or units from different data sources are joined together into a single file using non-unique identifiers, such as names, date of birth, addresses and other characteristics. **Synonyms: data matching, data linkage, entity resolution**.

Remote sensing

Acquisition of information about an object or phenomenon from a distant point.

S

Sample

A subset of the units of a population.

Sample survey

Survey for which the information is collected for some units of the target population only.

Sampling error

Difference between the estimate derived from a sample survey and the true value that would result if a census of the whole population were taken under the same conditions.

Sampling variation

Average of the squared differences between an estimate and the average of the estimates across all possible samples.

Secondary source information

Data from a secondary source was collected for a purpose other than producing statistical information.

Semi-interquartile range

Half the value of the interquartile range.

Spreadsheet

A software application that displays a table of cells arranged in rows and columns, in which the change of the contents of one cell can cause recalculation of other cells based on user-defined formulas.

Standard deviation

Square root of the variance.

Standard error

Square root of the sampling variance.

Statistical information

Data that have been recorded, classified, organized, related, or interpreted within a framework so that meaning emerges.

Statistical register

Data sets created for statistical purposes that are continuously updated with information about all units of a population.

Statistics

Type of information obtained through mathematical operations on data.

Structured data

Data that are organized into pre-defined items that each relates to a specific concept or data item.

Survey

Any activity to collect information in an organized and methodical manner about the characteristics of the units of a population. The word **survey** is often used to refer to a sample survey, as opposed to a census.

U

Unstructured data

Unstructured data are any data that are not arranged according to a pre-defined model.

Upper quartile

Value under which 75% of data points are found when arranged in increasing order. **Synonym: Third quartile**.

V

Variable

Characteristic that can be measured and that can assume different values.

Variance

Average of the squared differences between each data point and the centre of the distribution, measured using the mean.

W

Web scraping

The process through which information is gathered and copied from the web for further analysis.