Catalogue no. 12-004-X
ISBN 978-0-660-26572-8

Statistics: Power from Data!

Release date: April 3, 2001

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 STATCAN.infostats-infostats.STATCAN@canada.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.

Standard table symbols

The following symbols are used in Statistics Canada publications:

. not available for any reference period
.. not available for a specific reference period
... not applicable
0 true zero or a value rounded to zero
0^s value rounded to 0 (zero) where there is a meaningful distinction between true zero and the value that was rounded
^{p} preliminary
^{r} revised
x suppressed to meet the confidentiality requirements of the *Statistics Act*
^{E} use with caution
F too unreliable to be published
* significantly different from reference category ($p < 0.05$)

Published by authority of the Minister responsible for Statistics Canada

An HTML version is also available.

Cette publication est aussi disponible en français.

Acknowledgements

This electronic publication is the culmination of the efforts of many people.

We would particularly like to thank all of our colleagues in Dissemination, Communications, Methodology and Subject Matter divisions throughout Statistics Canada who assisted in reviewing this electronic publication.

Special thanks 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.

Joe Anne Legge
Research and Edit Officer
Statistics Canada

About *Statistics: Power from Data!*

Statistics: Power from Data! will assist readers in getting the most from statistics. Each chapter is intended to be complete in itself, allowing you to go directly to the topic you wish to learn more about without reading all of the other sections.

This web resource is published primarily for secondary students of Mathematics and Information Studies, although it will also be used by other students, teachers and the general population.

Statistics: Power from Data! was been created and modified using comments and requests from teachers, about the topics they would like to see covered, and the amount of time that could be devoted to them in a course.

Along with extensive text, this web resource contains exercises to help students consolidate their understanding of the material.

This resource aims to help students:

- gain confidence in using statistical information to complete study requirements
- 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.

The first section of this resource defines the basic concepts of "information, data and statistics" while the second explains how to plan the complete survey process. In the following sections, we detail each step of the process, such as sampling, data collection, graphing results, etc. Finally, we consider the use and misuse of statistics in society, the importance of confidentiality and the role of computers in producing statistics.

The length and amount of detail of a section does not reflect the importance of its topic to the overall survey process but rather, as mentioned above, how well it fits into the curriculum. 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 students but the details of data estimation are deemed too technical for this audience.

This product will be continually updated to keep information as relevant and topical as possible.

Inside the world of Statistics Canada

Statistics Canada has no written statement on its mission beyond what is said in the opening of the *Statistics Act*. Yet, a common understanding of the Agency's values is remarkably pervasive among its staff. The Agency's values are the guiding principles behind the work of all employees and the decisions of the Chief Statistician. Indeed, Dr. Fellegi, the former Chief Statistician, said it very clearly when he took office in September 1985:

> "I will continue to place emphasis on the foundation of the Agency's (Statistics Canada's) program, including vigilant attention to the relevance of our program and our service organization character; the principles of confidentiality, neutrality and scientific excellence, which I feel are absolutely crucial because they establish the intrinsic value of our information for users; on response burden; and co-ordination and integration."

Statistics Act states that our mandate is:

- to collect, compile, analyse, abstract and publish statistical information relating to the commercial, industrial, financial, social, economic and general activities and condition of the people;
- to collaborate with departments of government in the collection, compilation and publication of statistical information, including statistics derived from the activities of those departments;
- to take the Census of Population of Canada and the Census of Agriculture of Canada as provided in this Act;
- to promote the avoidance of duplication in the information collected by departments of government; and
- generally, to promote and develop integrated social and economic statistics pertaining to the whole of Canada and to each of the provinces thereof and to coordinate plans for the integration of those statistics.

Ensuring objectivity

By ensuring objectivity, Statistics Canada makes a fundamental contribution to the functioning of Canada's democratic system. The government and its political opposition can argue about their conflicting views on policy while agreeing on the underlying basic information. And the electorate can judge the performance of a government on the basis of its 'score card', a good portion of which is based on information compiled by Statistics Canada.

Protecting confidentiality

In collecting information from thousands of Canadian individuals and organizations, Statistics Canada has always put the highest priority on protecting the confidentiality of individual respondents' answers. With few exceptions, it is mandatory to respond to surveys collected under the *Statistics Act*. For its part, Statistics Canada is also obliged by law to ensure that individual answers are fully confidential. No other government agency, not even the Canadian Security and Intelligence Service or the Royal Canadian Mounted Police, is allowed access to individually identifiable responses.

Professionalism and reliability

Statistics Canada prides itself on the objectivity of the information it produces. But for its information to be considered authoritative, it must be reliable and its users must be persuaded that it was compiled in a thoroughly professional fashion. Statistics Canada has gone a long way in promoting reliability and professionalism. Many organizations across the world have given Statistics Canada the ranking of one of the top statistical agencies in the world.

Focusing on analysis

As Statistics Canada accentuated its analysis in the mid-1970s, the importance of this activity in disseminating data has grown. Analysis also plays a role in ensuring the relevance of the Agency's programs and priorities and improving communications with users. In the words of Canada's former Chief Statistician, Dr. Fellegi:

> "Analysis also popularizes our information. It highlights nuggets of information from among the mass of data we produce, giving Canadians information they can use and helping them know what the Agency does."

Reducing the response burden

Statistics Canada has always been conscious of the burden its surveys impose on respondents, particularly small businesses.

The Agency works at controlling the intrusiveness of its surveys by

- minimizing what we ask by using information that has already been collected through administrative purposes, such as tax or customs records
- shortening and simplifying questionnaires and reducing the number of surveys
- joint collection activities, such as joint federal-provincial surveys
- establishing guidelines for determining when surveys should be voluntary and when mandatory.

Data, information and statistics

Now that the world has entered the new millennium, we are facing new and challenging problems in our everyday lives. More now than ever before, governments, industry and society need reliable information to make better decisions in tackling these problems.

The need for an informed society is one reason why the Canadian education system is developing an emphasis on students gathering and processing data, and presenting the information in order to complete work requirements. However, before students can undertake such activities, it is important for them to have a sound understanding of the terms *data*, *information* and *statistics*.

Definitions

Data

Before one can present and interpret information, there has to be a process of gathering and sorting data. Just as trees are the raw material from which paper is produced, so too, can data be viewed as the raw material from which information is obtained.

In fact, a good definition of data is "**facts or figures from which conclusions can be drawn**".

Data, information and statistics are often misunderstood. They are actually different things, as Figure 1 shows.

Figure 1. Data collected on the weight of 20 individuals in your classroom

Data	Information	Statistics
20 kg, 25 kg	5 individuals in the 20-to-25-kg range	Mean weight = 22.5 kg
28 kg, 30 kg, etc.	15 individuals in the 26-to-30-kg range	Median weight = 28 kg

Data can take various forms, but are often numerical. As such, data can relate to an enormous variety of aspects, for example:

- the daily weight measurements of each individual in your classroom;
- the number of movie rentals per month for each household in your neighbourhood;
- the city's temperature (measured every hour) for a one-week period.

Other forms of data exist, such as radio signals, digitized images and laser patterns on compact discs.

Statistics Canada collects data every day to provide information.

Once data have been collected and processed, they are ready to be organized into information. Indeed, it is hard to imagine reasons for collecting data other than to provide information. This information leads to knowledge about issues, and helps individuals and groups make informed decisions.

In practice, informed decision-making can save countries millions of dollars (for example, through accurate targeting of government spending). It can also lead to life saving breakthroughs in medicine, and can help conserve the earth's natural environment.

Information

A good definition of information is "data that have been recorded, classified, organized, related, or interpreted within a framework so that meaning emerges".

Information, like data, can take various forms. Some examples of the different types of information that can be derived from data include:

- the number of persons in a group in each weight category (20 to 25 kg, 26 to 30 kg, etc.);
- the total number of households that did not rent a movie during the last month; and
- the number of days during the week where the temperature went above 20°C.

Some of the first known artifacts found in Canada provided ancient peoples with information that we still use today. The astrolabe is a scientific instrument dating back to 170 B.C. Complex planetary astrolabes were used to measure the altitude of the planets and stars to track their movements. From these calculations, latitude and time could also be determined.

The astrolabe

In the 16th century, a simpler nautical or mariner's astrolabe was developed for navigational use. The one pictured above is believed to be the astrolabe used by Christopher Columbus. He would have aligned the horizontal axis of the astrolabe with the horizon of the sky. He would then have directed the moveable pointer ("Alidade") towards the sun or polar star and read the position on its outer dial. The measurement noted was then compared to the astronomical tables in order to fix his latitude. In recent years, astronomers have used beams of radio waves to explore space. Also, the increase in television and video use has made visual information a popular tool.

In the 17th century, English philosopher Francis Bacon recognized the importance of knowledge. His quotation below is probably truer today than it has ever been, and is an important reason why the general public should have access to information.

> "Knowledge itself is power."
> — *Francis Bacon*

Statistics

Statistics represent a common method of presenting information. In general, statistics relate to numerical data, and can refer to the science of dealing with the numerical data itself. Above all, statistics aim to provide useful information by means of numbers.

Therefore, a good definition of statistics is "a type of information obtained through mathematical operations on numerical data".

Score of hockey game

Toronto Maple Leafs	Edmonton Oilers
6	5

Using the previous examples, some of the statistics that can be obtained include:

Statistics obtained

Information	Statistics
the number of persons in a group in each weight category (20 to 25 kg, 26 to 30 kg, etc.);	the average weight of students in your class
the total number of households that did not rent a movie during the last month; and	the minimum number of rentals your household had to make to be in the top 5% of renters for the last month; and
the number of days during the week where the temperature went above 20°C.	the minimum and maximum temperature observed each day of the week

A major role of Statistics Canada is to provide the Canadian community with statistics that will help society make informed decisions. Statistical information provided by Statistics Canada is used widely by governments, business people, doctors, farmers, teachers and students.

The provision of accurate and authoritative statistical information strengthens modern societies. It provides a basis for decisions to be made on such things as where to open schools and hospitals, how much money to spend on welfare payments and even which football players to replace at half-time! An example of statistical information that can be used for decision-making is given below.

"Average earnings of young people aged 15 to 24 have been declining since 1980. The impact of changes in overall economic activity on youth differs from that of other age groups. For example, during good economic times, job opportunities for young people grow, but their earnings do not appear to grow at the same rate as those of experienced workers. During 1970 to 1980 and 1985 to 1990, real average earnings of young people increased, but at a rate slightly lower than the overall rate of increase.

During the tough economic times between 1980 and 1985, and in the early 1990s, their average earnings fell about 20%, much more than the overall rate of decline.

In 1970 and 1980, the average earnings of young people aged 15 to 24 were around one-half of the overall national average. By 1995, they had declined to 31%. As a result of these changes, the position of young earners relative to other age groups consistently deteriorated between 1980 and 1995.

The average earnings for the 15 to 24 age group in 1980 was $13,191 while the average earnings in 1995 for the same age group was $8,199."

Source: *Statistics Canada, "The Daily", Tuesday, May 12, 1998.*

Examples of statistical information

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

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!

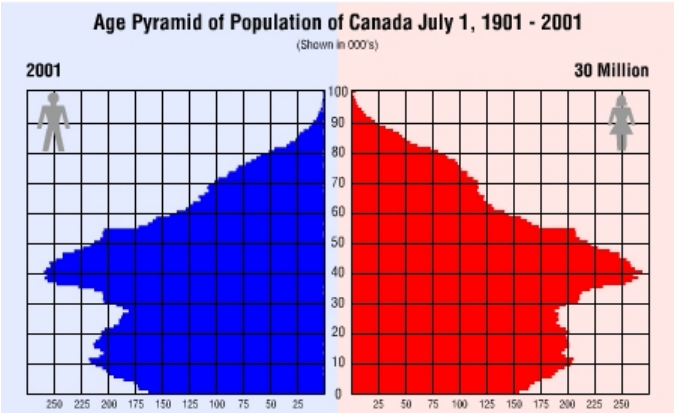
Selected occupations, Canada: 1911 Census

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**

Population pyramids

This is a chart of an age—sex pyramid from the 2001 annual population estimates. This chart shows statistical information on Canada's population by age group and sex. Canada's total population was 30,007,088 in 2001.

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



The Canadian Football League (CFL)

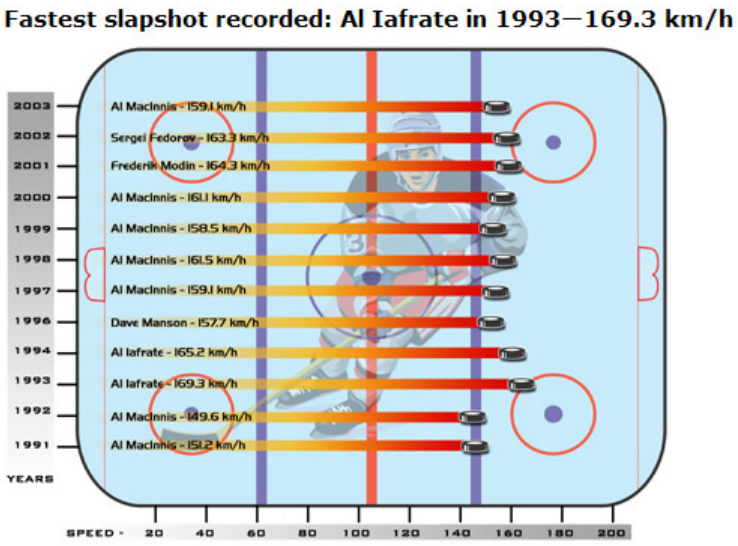
This table of statistical information shows how the Canadian football teams performed in the 2002 season for the total all-purpose yards standings in the Canadian Football League (CFL).

Official 2002 CFL statistics - team offence

Team	Yards	Rush	Pass
Montreal Alouettes	7,170	2,342	5,242
Winnipeg Blue Bombers	6,993	1,850	5,544
Calgary Stampeders	6,291	1,930	4,766
Edmonton Eskimos	6,237	2,061	4,648
B.C. Lions	6,230	2,183	4,357
Hamilton Tiger-Cats	6,139	1,558	4,807
Saskatchewan Roughriders	5,701	2,518	3,600
Ottawa Renegades	5,100	1,488	3,933
Toronto Argonauts	4,436	1,664	3,162

The National Hockey League (NHL)

This illustration of statistical information refers to the National Hockey League (NHL) players having the fastest slapshots since 1991. The illustration shows information about the speed of each player's slapshot.



Later on, you will see how easy it is to be misled by statistical information. If used wisely, statistics can be a powerful tool in decision-making.

Exercise

1. In your own words define the terms *data*, *information* and *statistics*. Give examples of each.
2. Put the following terms in correct logical order: Knowledge, Data, Information.
3. Identify three current political, economic or social issues for which information is necessary. Then, describe the information that is needed for each issue.
4. In all of the examples of statistical information illustrated in this chapter, which one contains the fewest observations and which one contains the most?
5. In the 1911 Census Occupations example, three occupations have no female members (0 observations). Does this mean that no information is available for females in that occupation?
6. In the Canadian Football League (CFL) example, which team recorded the lowest number of yards for yards on passes received during the 2002 season?
7. In the National Hockey League (NHL) example, how many units of information (hockey players) are illustrated?
8. Who do you think might need the information in the 1911 Census Occupations example?
9. Who do you think might need the information in the Population Pyramids example, and for what purpose?
10. Does the information in the CFL example accurately show, with regard to total all purpose yards, which team performed better than others for the 2002 season? Explain.
11. Which example required use of a scientific instrument to collect the data?
12. Which example shows all of the individual observations collected?

Class activity

1. The examples from this chapter also illustrate the variety of ways in which statistics can be presented. Look in newspapers or journals for other ways statistics are shown. Be careful to distinguish the difference between data, information and statistics, as in Exercise 1.
2. ARCHIVED - Counting Canadians

Answers

1.

2. The correct logical order is: Data, Information, Knowledge.

3.

4. The example with the fewest observations is the Canadian Football League table with 8 observations. Population pyramids presents the most number of observations.

5. The three occupations that do not have any female units are: Undertakers, Coachmen and Grooms, and Sailors and Seamen. However, the value 0 can still be regarded as information.

6. The Toronto Argonauts has the lowest number of yards (4,436) in the Canadian Football League for yards gained on passes received in the 2002 season.

7. There is information on twelve observations on 5 hockey players in the NHL example.

8. History researchers and history students might be interested in the information from the 1911 Census Occupations example.

9. Governments planning health and social policies might be interested in the information in the Population pyramids example.

10. No, statistics on the number of total all-purpose yards does not measure a team's overall performance.

11. The NHL example, the judges required the use of a radar gun to measure the speed of the slapshots.

12. None of the examples in this chapter show all of the individual observations collected. (For instance, in the NHL example, each player made more than one slapshot, but only the fastest shot is shown.)

Planning a survey

On 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 that address the following issues:

Survey objectives

A survey plan begins with objectives that describe why and for whom the survey is being done. The survey objectives tell a lot about the data that need to be collected. The objectives also help determine the population to be targeted.

> For 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 make some more refined 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 organising committee.

The survey plan will show 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 analysed.

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 (the population for which information is required) and the survey population (the population actually completing the survey) differ for practical reasons, even though they should, in reality, be the same.

> 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 religious seminaries or military schools. These types of graduates would be reasonably assured of securing employment in their respective fields. 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 minimise the travel involved.

Data requirements

To determine what kind 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?
    - 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?
- Concerning security, 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 that 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 analyse 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. Respondent burden is affected by

- the number of questions asked
- the intrusiveness of the questions
- the number of times the respondent is contacted (for a 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 accuracy pursued and the resources available will determine the choice among three main types of data collection.

1. A **census** is a survey that collects information from all the people in a group or population
2. A **sample survey** collects information from only a part (a sample) of a population. It is possible to estimate results for a total population using data that is collected from a sample.
3. **Administrative data** is collected through an organisation and is used as an alternative to a survey.

Each has advantages and inconveniences and the choice of collection type will depend on various factors. See Types of data collection.

In our example, the organising 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 reasonable 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 permits 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 operation, it reduces the data collection and processing time, while allowing more time for planning.

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 a certain bias, 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 larger the sample size), the smaller the sampling error will be. Also, it is possible to estimate the sampling error associated with a particular sampling plan, and try to minimize it. See Sampling error.

By choosing to do a census, you can avoid errors related to sample variation, but all surveys also risk having sources of "non-sampling error". 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 can be avoided as much as possible by paying attention to quality control throughout every step of the survey process. See Non-sampling error.

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 a lot of variability in the population, the sample size will need to be bigger to obtain the specified level of reliability. See Sample size.

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 "empty" tables help you verify whether the questions you are considering will allow you to reach your 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's design is based on the survey's data requirements and analysis plan. As you 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.

It's important to ensure that the questions relate to the survey objectives and that each question is relevant. See Questionnaire design.

Data collection methods

Planning the method of data collection is an important step: you will need to consider the costs, physical resources, and time required to conduct the survey.

Select the best method to gather the required data. Keep in mind that cost of the survey and data quality will be directly impacted by the method that you choose. There are several options available: the personal interview (face-to-face or by telephone, with or without computer assistance) and the self-completed questionnaire.

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 any help from an interviewer. It can be returned by mail or through a drop-off system or completed online. See Data collection methods.

> 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 processing the questionnaire responses into output. The tasks involved in data processing include: coding, data capture, editing, dealing with invalid or missing data and, if necessary, creating derived variables. In short, the aim in this step is to produce a file of data that is as free of errors as possible. See Data processing.

Quality control

This process identifies errors and verifies results. No matter how much planning and testing goes into a survey, something unexpected will often 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, data editing, 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
- describing what characterizes the data using measures of central tendency and measures of spread
- displaying the data through different graph types
- 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 what students told them.

Exercises

1. What are the three first steps in planning a survey?

2. What type of data collection (census, sample survey or use of administrative data) would be 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?

3. What can be done to reduce error in a survey as much as possible?

Class activities

- Choose a topic that you would like to investigate concerning your school, family or neighbourhood. Express your topic as a clear question that you would like to have answered. Then draw up a plan for how you would collect and analyse the information required to answer that question.

- Think of a possible class survey project that would collect useful information for a committee or group in your school or community.

Answers

1.

a. Define the survey objectives – Why is this survey being conducted?
b. Define the target population – Whom will the collected information be about?
c. Define the data requirements – What do I need to know?

2.

a. Administrative data: the information would be provided by doctors when they fill in death certificates.
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. You can omit sampling error completely by doing a census or minimize it by increasing the size of the sample.

You can minimize non-sampling errors by paying attention to quality control throughout the survey process.

Sampling methods

The need for accurate information in order to make informed decisions is emphasized throughout this online publication (see the section on Data, information and statistics). Statistics are an important type of information and statistical agencies play an important part in producing such information. In order to do this, the agencies conduct censuses and sample surveys and use administrative records. These three types of data collection, along with their advantages and disadvantages, are explained in the Data collection section.

The present section focuses on sampling and on estimation (which is the process of taking information gathered from a sample and extending it to the whole population).

Selection of a sample

Sampling allows statisticians to draw conclusions about a whole by examining a part. It enables us to estimate 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 survey—and how this information can be applied to the entire population.

It is essential that a sample survey be correctly defined and organized. If the wrong questions are posed to the wrong people, statisticians will not receive information that will be useful when applied to the entire population.

In the context of a national statistical agency like Statistics Canada, the following steps are needed to select a sample and ensure that this sample will fulfill its goals.

Establish the survey's objectives

The first step in planning a useful and efficient survey is to specify the objectives with as much detail as possible. Without objectives, the survey is unlikely to generate usable results. Clarifying the aims of the survey is critical to its ultimate success. The initial users and uses of the data should be identified at this stage.

The pros and cons of a census versus a sample survey or the use of administrative records should be evaluated and a decision made as to the most appropriate method. (At this point, we will assume that a sample survey is the best way to proceed in order to obtain the information we need. This assumption will hold true for the remainder of the sample selection steps, even though many of the steps mentioned will also apply to the other methods.)

Define the target population

The target population is the total population for which the information is required. For example, if you were to conduct a survey about the most popular types of cars in Saskatchewan, then the target population would be every car in Saskatchewan. The units that make up the population must be described in terms of characteristics that clearly identify them. Specifically, the target population is defined by the following characteristics:

- Nature of data required: about 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 time period covered by the survey.
- Other characteristics, such as socio-demographic characteristics (interest in a particular age group, for example) or type of industry.

Decide on the data to be collected

he data requirements of the survey must be established. To ensure that the requirements are operationally sound, the necessary data terms and definitions also need to be determined.

Set the level of precision

As mentioned in the section on Sampling error, there is a level of uncertainty associated with estimates coming from a sample. For example, if you are trying to estimate 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 who the 5 sampled students are. If the 5 sampled students also live close to the school, the results will not be able to represent the class accurately. This sample-to-sample variation is what causes the sampling error. Statisticians can estimate the sampling error associated with a particular sampling plan, and try to minimize it.

When designing a survey, the acceptable level of uncertainty in the survey estimates has to be established. This level depends on what the end use of the results will be and on the size of the overall budget. The bigger the budget, the more resources available, and thus, less chance for error. And if the end result is to serve a specific purpose, then the acceptable level of uncertainty would be smaller than an end result that is simply looking for general trends.

The level of uncertainty will also be determined by the sample size. Increasing the sample size will decrease the sampling error. (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 possible samples.)

The sample design

Once the objectives, guidelines and definitions have been worked out, the statistician can work on the survey plan. The survey plan is divided into three parts:

- Sample design: how the sample will be collected.
- Estimation techniques: how the results from the sample will be extended to the whole population.
- Measures of precision: how the sampling error will be measured.

The estimation techniques and measures of precision are discussed in a later section. For the moment, we will look at the sample design. The following steps lead to the complete 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.

The survey population

The target population must be defined early in the survey-designing process. This is the population for which information is required. However, some members of the population have to be excluded because of operational constraints: the high cost of collecting data in some remote areas, the difficulty of identifying and contacting certain components of the target population, etc. For example, because it would be too difficult to locate and survey each car owned by every resident in Saskatchewan, a survey population of just the major cities and towns might be conducted instead. When some of the members of the target population are excluded, we call the included population the survey population or, what is sometimes called, an *observed population*. The target population is the population we **want to observe** while the survey population is the population we **can observe**.

The goal of this process is to have the survey population as close as possible to the target population. It is also very important that the users of the data be informed 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 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 names and addresses that provide direct access to 'individuals' (e.g., a list of hospitals, a list of restaurants, a list of students at a university). Area frames are a list of geographic areas that provide indirect access to individuals (e.g., the neighbourhoods in a city). This type of access is called indirect because first, a list of geographic areas must be selected and then, access to individuals within each selected area must be worked out.

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.

When there is no single frame that is appropriate, multiple frames can be used. Some sampling techniques using both types of frames will be discussed later.

A good frame should be complete and up-to-date; no member of the survey population should be excluded from the frame or duplicated on the frame (represented more than once); and no unit that is not part of the population (e.g., deceased persons) should be on the frame. The frame chosen 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.

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 provides the information needed by the survey.
- The unit of reference or unit of analysis—the unit about which information is provided—is used to analyse 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 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, such as budget and time. 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 would 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 level, 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. You will need a bigger sample if your method is not the most efficient. But because of operational constraints and the unavailability of an adequate frame, you cannot always use the most efficient technique.

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 'chance' of being selected, and that chance can be quantified. This is not true for non-probability sampling; every item in a population does

not have an equal chance of being selected. The next section will describe features of both types of sampling and detail some of the methods related to each type.

Probability sampling

Probability sampling involves the selection of a sample from a population, based on the principle of randomization 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 probability of inclusion can be calculated, reliable estimates can be produced along with estimates of the sampling error, and inferences can be made about the population.

There are several different ways in which a probability sample can be selected. The method chosen depends on a number of factors, such as the available sampling frame, how spread out the population is, how costly it is to survey members of the population and how users will analyse the data. When choosing a probability sample design, your goal should be 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.

The following are the most common probability sampling methods:

- simple random sampling
- systematic sampling
- sampling with probability proportional to size
- stratified sampling
- cluster sampling
- multi-stage sampling
- multi-phase sampling

Simple random sampling

In *simple random sampling*, each member of a population has an equal chance of being included in the sample. Also, each combination of members of the population has an equal chance of composing the sample. Those two properties are what defines simple random sampling. 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 were 2,000, then 2,000 numbers between 1 and 10,000 would need to be randomly generated by a computer. Each number will have the same chance of being generated by the computer (in order to fill the simple random sampling requirement of an equal chance for every unit). The 2,000 telephone entries corresponding to the 2,000 computer-generated random numbers would make up the sample.

Simple random sampling can be done with or without replacement. A sample with replacement means that there is a possibility that the sampled telephone entry may be selected twice or more. Usually, the simple random sampling approach is conducted without replacement because it is more convenient and gives more precise results. For the purpose of these descriptions, when we discuss simple random sampling, we will refer to sampling without replacement.

Simple random sampling is the easiest method of sampling and it is the most commonly used. Advantages of this technique are that it does not require any additional information on the frame (such as geographic areas) other than the complete list of members of the survey population along with information for contact. Also, since simple random sampling 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 makes no use of auxiliary information present on the frame (i.e., number of employees in each business) that could make the design of the sample more efficient. And although it is easy to apply simple random sampling to small populations, it can be expensive and unfeasible for large populations because all elements must be identified and labeled prior to sampling. It can also be expensive if personal interviewers are required since the sample may be geographically spread out across the population.

A lottery draw is a good example of simple random sampling. For example, when a sample of 6 numbers is randomly generated from a population of 49, each number has an equal chance of being selected and each combination of 6 numbers has the same chance of being the winning combination. Even though people tend to avoid combinations such as 1-2-3-4-5-6, it has the same chance of being the winning set of numbers as the combination of 8-15-21-28-32-40.

Example 2: Suppose your school has 500 students and you need to conduct a short survey on the quality of the food served in the cafeteria. You decide that a sample of 10 students should be sufficient for your purposes. In order to get your sample, you assign a number from 1 to 500 to each student in your school. To select the sample, you use a table of randomly generated numbers. All you have to do is pick a starting point in the table (a row and column number) and look at the random numbers that appear there. In this case, since the data run into three digits, the random numbers would need to contain three digits as well. Ignore all random numbers after 500 because they do not correspond to any of the students in the school. Remember that the sample is without replacement, so if a number recurs, skip over it and use the next random number. The first 10 different numbers between 001 and 500 make up your sample.

Example 3: 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 asking them which of the listed movies are their favourites. To create the list of movies needed for your survey, you decide to sample 100 of the 1,000 best horror movies of all time. The horror movie population is divided evenly into classic movies (those filmed in or before 1969) and modern movies (those filmed in or later than 1970). One way of getting a sample would be to write out all of the movie titles on slips of paper and place them in an empty box. Then, draw out 100 titles and you will have your sample. By using this approach, you will have ensured that each movie had an equal chance of selection

You can also calculate the probability of a given movie being selected. Since we know the sample size (**n**) and the total population (**N**), calculating the probability of being included in the sample becomes a simple matter of division:

Probability of being selected (same for each horror movie)

= (**n** ÷ **N**) x 100%

= (100 ÷ 1,000) x 100%

= 10%

This means that every movie title on the list has a 10% or a 1 in 10 chance of being selected.

You can see that that one disadvantage of simple random sampling (not the only disadvantage, but an important one) is that even if you know that the population is made up of 500 classic movies and 500 modern movies and you know each movie's release date from the sampling frame, no use is made of this information. This sample might contain 77 classic movies and 23 modern movies, which would not be representative of the whole horror movie population.

There are ways to overcome this problem (these will be briefly discussed in the Estimation section), but there are also ways to account for this information. (This will also be discussed later, under the section on Stratified sampling.)

Systematic sampling

Sometimes called *interval sampling*, *systematic sampling* means that there is a gap, or interval, between each selected unit in the sample. In order to select a systematic sample, you need to 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 would be the first number included in your sample. Using the sample above, you would select a number between 1 and 4 from a table of random numbers. 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 **K**$^{\textbf{th}}$ (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).

Using the example above, you can see that with a systematic sample approach 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 sampling of 100 units was selected. The main difference is that with simple random sampling, any combination of 100 units would have a chance of making up the sample, while with systematic sampling, there are only four possible samples. From that, we can see how precise systematic sampling is compared with simple random sampling. The population's 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.

Example 4: Imagine you have to conduct a survey on student housing for your university or college. Your school has an enrolment of 10,000 students and you want to take a systematic sample of 500 students. In order to do this, you must first determine what your sampling interval (**K**) would be:

Total population ÷ sample size **=** sampling interval
N ÷ **n** **=** **K**
= 10,000 ÷ 500
= 20

To begin this systematic sample, all students would have to be assigned sequential numbers. The starting point would be chosen by selecting a random number between 1 and 20. If this number were 9, then the 9th student on the list would be selected along with every 20th student thereafter. The sample of students would be those corresponding to student numbers 9, 29, 49, 69...9,929, 9,949, 9,969 and 9,989.

In the examples used thus far, the sampling interval **K** was always a whole number, but this is not always the case. For example, if you want a sample of 30 from a population of 740, your sampling interval (or **K**) will be 24.7. In these cases, there are a few options to make the number easier to work with. You can round the number—either round it up to the nearest whole number or round it down. Rounding down will ensure that you select at least the number of units you originally wanted (and you can then delete some units to get the exact sample size you wanted). Techniques exist to adapt systematic sampling to the

case where **N** (total population) is not a multiple of **n** (sample size), but still give a sample exactly the same as the **n** units. These techniques will not be discussed here.

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) and that the sample is distributed evenly over the listed population. The biggest drawback of the systematic sampling method is that if there is some cycle in the way the population is arranged on a list and if that cycle 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 5: 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 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 have a chance of being included in the sample, but it does not require that this chance 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 it has 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

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 (and others that exist) can be used to sample within each stratum. The sampling method can vary from one stratum to another. When simple random sampling is used to select the sample within each stratum, the sample design is called *stratified simple random sampling*. A population can be stratified by any variable that is available for all units on the sampling frame prior to sampling (e.g., age, sex, province of residence, income, etc.).

Why do we need to create strata? There are many reasons, the main one being that it can make the sampling strategy more efficient. It was mentioned earlier that you need a larger sample to get a more accurate estimation of a characteristic that varies greatly from one unit to the other than for a characteristic that does not. 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 (e.g., income) and are considerably different from units in other strata (e.g., occupation, type of dwelling) 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 simple random sampling approach in the whole population without stratification, the sample would need to be larger than the total of all stratum samples to get an estimate of total income with the same level of precision.

Stratified sampling ensures an adequate sample size for sub-groups in the population of interest. When a population is stratified, each stratum becomes an independent population and you will need to decide the sample size for each stratum.

Example 6: Suppose you want to estimate how many high school students have part-time jobs at the national level and also in each province. 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 on average with just a little over 100 people from Prince Edward Island, since they account for less than half of a percent of the whole Canadian population. This sample would probably not be large enough for the kind of detailed analysis you had in mind. Stratifying your list by province, again assuming that this information is available, and then selecting a sample size for each province would allow you to decide on the exact sample size needed for that specific province. Thus, in order to get good representation of Prince Edward Island, you would use a larger sample than the one allotted to it by the simple random sampling approach.

Example 7: An Ontario school board wanted to assess student opinion on dropping Grade 13 from the secondary school program. They decided to survey students from Elmsview High School. To ensure a representative sample of students from all grade levels, the school board used a stratified sampling technique.

In this case, the strata were the five grade levels (grades 9 to 13). The school board then selected a sample within each stratum. The students selected in this sample were extracted using simple random or systematic sampling, making up a total sample of 100 students.

Stratification is most useful when the stratifying variables are

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

Cluster sampling

Sometimes it is too expensive to spread a sample across the population as a whole. 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 group.

Examples of clusters are factories, schools and geographic areas such as electoral subdivisions. The selected clusters are used to represent the population.

Example 8: Suppose you are a representative from an athletic organization wishing to find out which sports Grade 11 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 schools provide clusters of samples. Then every Grade 11 student in all 100 clusters is surveyed. In effect, the students in these clusters represent all Grade 11 students in Canada.

Example 9: Imagine that the municipal council of a small city wants to investigate the use of health care services by residents.

First, the council requests from Statistics Canada electoral subdivision maps that identify and label each city block. From these maps, the council creates a list of all city blocks. This list will serve as the sampling frame.

Every household in that city belongs to a city block, and each city block represents a cluster of households. The council randomly picks a number of city blocks. Using the simple random sample approach, then the council creates a list of all households in the selected city blocks; these households make up the survey sample.

As mentioned, cost reduction is a reason for using cluster sampling. It creates 'pockets' of sampled units instead of spreading the sample over the whole territory. Another reason is that sometimes a list of all units in the population (a requirement when conducting simple random sample, systematic sample or sampling with probability proportional to size) is not available, while a list of all clusters is either available or easy to create.

In most cases, the main drawback is a loss of efficiency when compared with simple random sampling. It is usually better to survey a large number of small clusters instead of a small number of large clusters. This is 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 two previous examples, students in the same school tend to participate in the same types of sports (depending on the facilities available at their school); similarly, elderly people have a tendency to live in specific neighbourhoods and to be heavy users of health services.

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 city blocks do not all have the same number of households, and you must interview every student or household in your sample, the final size may be larger or smaller than you expected.

Multi-stage sampling

Multi-stage sampling is like the cluster method, except that it involves picking a sample from within each chosen cluster, rather than including all units in the cluster. This type of sampling requires at least two stages. In the first stage, large groups or clusters are identified and selected. These clusters contain more population units than are needed for the final sample.

In the second stage, population units are picked from within the selected clusters (using any of the possible probability sampling methods) for a final sample. If more than two stages are used, the process of choosing population units within clusters continues until there is a final sample.

Example 10: In Example 8, a cluster sample would choose 100 schools and then interview every Grade 11 student from those schools. Instead in multi-stage sampling, you could select more schools, get a list of all Grade 11 students from these selected schools and select a random sample (e.g., simple random sample) of students from each school. This would be a two-stage sampling design.

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 class. This would be a three-stage sampling design. Each time we add a stage, 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 Grade 11 students). If you wanted a bigger sample, you could select schools with more students; and 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*.

With a three-stage design, you could select a sample of 400 schools, then select two Grade 11 classes per school (assuming that there are two or more Grade 11 classes per school). Finally, you could 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 will be more spread out.

You can see from this example that with multi-stage sampling, you still have the benefit of a more concentrated sample for cost reduction. However, the sample is not as concentrated as other clusters and the sample size is still bigger than for a simple random sample size. Also, you do not need to have a list of all of the students in the population. All you need is a list of the classes from the 400 schools and a list of the students from the 800 classes. Admittedly, more information is needed in this type of sample than what is required in cluster sampling. However, multi-stage sampling still saves a great amount of time and effort by not having to create a list of all of the units in a population.

Multi-phase sampling

A *multi-phase sample* collects basic information from a large sample of units and then, for a subsample of these units, collects more detailed information. 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 multi-stage sampling, despite the similarities in name. Although multi-phase sampling also involves taking two or more samples, all samples are drawn from the same frame and at each phase the units are structurally the same. However, as with multi-stage sampling, the more phases used, the more complex the sample design and estimation will become.

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

Example 11: 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 could be conducted whose only question is "Is part or all of your farm devoted to cattle farming?" 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).

Multi-phase sampling can be used when there is insufficient budget to collect information from the whole sample, or when doing so would create excessive burden on the respondent, or even when there are very different costs of collection for different questions on a survey.

Example 12: A health survey asks participants 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. This second sample will take part in the medical tests.

Non-probability sampling

The difference between *probability* and *non-probability sampling* has to do with a basic assumption about the nature of the population under study. In probability sampling, every item has a chance of being selected. In non-probability sampling, there is an assumption that there is an even distribution of characteristics within the population. This is what makes the researcher believe that any sample would be representative and because of that, results will be accurate. For probability sampling, randomization is a feature of the selection process, rather than an assumption about the structure of the population.

In non-probability sampling, 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.

Reliability cannot be measured in non-probability sampling; the only way to address data quality is to compare some of the survey results with available information about the population. Still, there is no assurance that the estimates will meet an acceptable level of error. Statisticians are reluctant to use these methods because there is no way to measure the precision of the resulting sample.

Despite these drawbacks, non-probability sampling methods can be useful when descriptive comments about the sample itself are desired. Secondly, they are quick, inexpensive and convenient. There are also other circumstances, such as in applied social research, when it is unfeasible or impractical to conduct probability sampling. Statistics Canada uses probability sampling for almost all of its surveys, but uses non-probability sampling for questionnaire testing and some preliminary studies during the development stage of a survey.

Most non-sampling methods require some effort and organization to complete, but others, like convenience sampling, are done casually and do not need a formal plan of action.

The most common types are listed below:

- convenience or haphazard sampling
- volunteer sampling
- judgement sampling
- quota sampling

Convenience or haphazard sampling

Convenience sampling is sometimes referred to as *haphazard* or *accidental sampling*. It is not normally representative of the target population because sample units are only selected if they can be accessed easily and conveniently.

There are times when the average person uses convenience sampling. A food critic, for example, may try several appetizers or entrees to judge the quality and variety of a menu. And television reporters often seek so-called 'people-on-the-street interviews' to find out how people view an issue. In both these examples, the sample is chosen randomly, without use of a specific survey method.

The obvious advantage is that the method is easy to use, but that advantage is greatly offset by the presence of bias. Although useful applications of the technique are limited, it can deliver accurate results when the population is homogeneous.

For example, a scientist could use this method to determine whether a lake is polluted. Assuming that the lake water is well-mixed, any sample would yield similar information. A scientist could safely draw water anywhere on the lake without fretting about whether or not the sample is representative.

Examples of convenience sampling include:

- the female moviegoers sitting in the first row of a movie theatre
- the first 100 customers to enter a department store
- the first three callers in a radio contest.

Volunteer sampling

As the term implies, this type of sampling occurs when people volunteer their services for the study. In psychological experiments or pharmaceutical trials (drug testing), for example, it would be difficult and unethical to enlist random participants from the general public. In these instances, the sample is taken from a group of volunteers. Sometimes, the researcher offers payment to entice respondents. In exchange, the volunteers accept the possibility of a lengthy, demanding or sometimes unpleasant process.

Sampling voluntary participants as opposed to the general population may introduce strong biases. Often in opinion polling, 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 large selection bias.

Television and radio media often use call-in polls to informally query an audience on their views. The Much Music television channel uses this kind of survey in their CombatZone program. The program asks viewers to cast a vote for one of two music videos by telephone, e-mail or through their online website.

Oftentimes, there is no limit imposed on the frequency or number of calls one respondent can make. So, unfortunately, a person might be able to vote repeatedly. It should also be noted that the people who contribute to these surveys might have different views than those who do not.

Judgement sampling

This approach is used when a sample is taken based on certain judgements about the overall population. The underlying assumption is that the investigator will select units that are characteristic of the population. The critical issue here is objectivity: how much can judgment be relied upon to arrive at a typical

sample? *Judgement sampling* is subject to the researcher's biases and is perhaps even more biased than haphazard sampling. Since any preconceptions the researcher may have are reflected in the sample, large biases can be introduced if these preconceptions are inaccurate.

Statisticians often use this method in exploratory studies like pre-testing of questionnaires and focus groups. They also prefer to use this method in laboratory settings where the choice of experimental subjects (i.e., animal, human, vegetable) reflects the investigator's pre-existing beliefs about the population.

One advantage of judgement sampling is the reduced cost and time involved in acquiring the sample.

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 sub-populations have been selected. Since there are no rules as to how these quotas are to be filled, *quota sampling* is really a means for satisfying sample size objectives for certain sub-populations.

The quotas may be based on population proportions. For example, if there are 100 men and 100 women in a population and a sample of 20 are to be drawn to participate in a cola taste challenge, you may want to divide the sample evenly between the sexes—10 men and 10 women. Quota sampling can be considered preferable to other forms of non-probability sampling (e.g., judgement sampling) because it forces the inclusion of members of different sub-populations.

Quota sampling is somewhat similar to stratified 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 it is usually left up to the interviewer to decide who is sampled. This results in selection bias. Thus, quota sampling is often used by market researchers (particularly for telephone surveys) instead of stratified sampling, because it is relatively inexpensive and easy to administer and has the desirable property of satisfying population proportions. However, it disguises potentially significant bias.

As with all other non-probability sampling methods, 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.

Example 1: The student council at Cedar Valley Public School wants to gauge student opinion on the quality of their extracurricular activities. They decide to survey 100 of 1,000 students using the grade levels (7 to 12) as the sub-population.

The table below gives the number of students in each grade level.

Table 1. Number of students enrolled at Cedar Valley Public School, by grade

Grade level	Number of students	Percentage of students (%)	Quota of students in sample of 100
7	150	15	15
8	220	22	22
9	160	16	16
10	150	15	15
11	200	20	20
12	120	12	12
Total	**1,000**	**100**	**100**

The student council wants to make sure that the percentage of students in each grade level is reflected in the sample. The formula is:

Percentage of students in Grade 10
= (number of students ÷ number of students) x 100%
= (150 ÷ 1,000) x 100
= 15%

Since 15% of the school population is in Grade 10, 15% of the sample should contain Grade 10 students. Therefore, use the following formula to calculate the number of Grade 10 students that should be included in the sample:

Sample of Grade 10 students
= (15% of 100) x 100
= 0.15 x 100
= 15 students

The main difference between stratified sampling and quota sampling is that stratified sampling would select the students using a probability sampling method such as simple random sampling or systematic sampling. In quota sampling, no such technique is used. The 15 students might be selected by choosing the first 15 Grade 10 students to enter school on a certain day, or by choosing 15 students from the first two rows of a particular classroom. Keep in mind that those students who arrive late or sit at the back of the class may hold different opinions from those who arrived earlier or sat in the front.

The main argument against quota sampling is that it does not meet the basic requirement of randomness. Some units may have no chance of selection or the chance of selection may be unknown. Therefore, the sample may be biased.

It is common, but not necessary, for quota samples to use random selection procedures at the beginning stages, much in the same way as probability sampling does. For instance, the first step in multi-stage sampling would be randomly selecting the geographic areas. The difference is in the selection of the units in the final stages of the process.

In multi-stage sampling, units are based on up-to-date lists for selected areas and a sample is selected according to a random process. In quota sampling, by contrast, each interviewer is instructed on how many of the respondents should be men and how many should be women, as well as how many people should represent the various age groups. The quotas are therefore calculated from available data for the population, so that the sexes, age groups or other demographic variables are represented in the correct proportions. But within each quota, interviewers may fail to secure a representative sample of respondents. For example, suppose that an organization wants to find out information about the occupations of men aged 20 to 25. An interviewer goes to a university campus and selects the first 50 men aged 20 to 25 that she comes across and who agree to participate in her organization's survey. However, this sample does not mean that these 50 men are representative of all men aged 20 to 25.

Quota sampling is generally less expensive than random sampling. It is also easy to administer, especially considering the tasks of listing the whole population, randomly selecting the sample and following-up on non-respondents can be omitted from the procedure. Quota sampling is an effective sampling method when information is urgently required and can be carried out independent of existing sampling frames. In many cases where the population has no suitable frame, quota sampling may be the only appropriate sampling method.

Estimation

As we now know, the goal of conducting surveys is to obtain information about a particular population. When the sample has been selected and the information collected (see the Data collection chapter) and processed (see the Data processing chapter), there still remains the task of linking the information gathered from the sample back to the overall population.

Estimation is the process of determining a likely value for a variable in the survey population, based on information collected from the sample. 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: estimates for 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; or the total number of cars possessed by the whole survey population.

Underpinning the estimation process is the sampling weight of a unit, which indicates the number of units in the population (including the sampling weight) that are represented by this sampled unit. The sampling weight is the inverse of the unit's probability of selection.

- **Example 1:** Suppose that the city of Winnipeg has decided to award bus travellers with free one-year bus passes as a way of promoting its services. A simple random sample of 10 people is selected from the 30 passengers on a city bus. Since simple random sampling gives equal probability of selection to every member of the population (in this case, all passengers on the bus), each passenger had one chance out of three of being selected. This translates into a sampling weight of three for every selected unit. This means that each person in the sample represents three persons in the population —himself or herself, plus two other persons

  To estimate this sampling weight, one could take the survey information for the 10 selected passengers and copy it three times to create an artificial population of 30. Totals, averages or proportions for the real population could then be estimated by the corresponding statistics computed using the artificial population. Instead of doing this, however, survey statisticians attach a sampling weight to each unit in the sample and take this weight into account when estimating.

  If one person in a sample (with a sampling weight of 18) had blue eyes and brown hair, then it is as if a total of 18 people in the population had blue eyes and brown hair.

  Example 2: You are conducting a survey to determine the total number of people living on your street and the average number of cars owned by each household. You decide to select a systematic sample of 5 households from the 20 on your street and intend to use that sample to estimate the totals you are looking for. The following table summarizes the information that you gathered during your interviews with the sampled households:

Table 1. Sample of households on Redwood Street

Household number	Number of persons	Number of cars	Probability of selection	Sampling weight
1	1	0	1/4	4
2	4	2	1/4	4
3	2	1	1/4	4
4	2	1	1/4	4
5	3	2	1/4	4

- The selection probability of 1 in 4 comes from the fact that systematic sampling gives an equal chance of being selected to each household on your street. The sampling weight of 4 is just the inverse of that probability. When estimating, you have to look at the characteristics of each sampled household. In this case, it is decided that 4 households from the population of 20 on your street have the same characteristics.

  In order to estimate the total number of persons living on your street, you have to multiply the number of persons in a household by the number of households in that sampling weight, then add up all the final numbers. For example, there are 4 one-person households (represented by Household number 1), 4 four-person households, 8 two-person households (four households represented by Household number 3 and four households represented by Household number 4) and 4 three-person households. The estimation of the total number of persons would then be:

  Estimated number of persons living on your street
  = (4 x 1) + (4 x 4) + (8 x 2) + (4 x 3)
  = 48 people

  To estimate the average number of cars per household, you proceed in the same manner. Get an estimate of the total number of cars owned by households on your street and then, divide the estimate by the actual number of households on the street. For example, there are 4 households without a car (represented by Household number 1), 8 households with two cars (represented by Household number 2 and Household number 5), 8 households with one car each (represented by Household number 3 and Household number 4).

Estimated number of cars
= (4 x 0) + (8 x 2) + (8 x 1)
= 24 cars

Estimated average
= 24 ÷ 20
= 1.2 cars per household

Self-weighting designs

It is not always the case that all sampled units had the same sampling weight. Some designs give unequal probability of selection to units, resulting in units within the same sample having different sampling weights. Answers from one household or business could represent the answers for 200 units of the population, while the answers from another could represent only 50 units in the population.

When every unit in the sample has the same sampling weight, the sampling design is said to be self-weighted. This kind of design is time-saving and operationally convenient, particularly for large samples. Because every unit has the same weight, those weights can be ignored when estimating averages and proportions. The average for the sample gives an appropriate estimate of the average for the whole population.

Simple random sampling and systematic sampling are examples of self-weighted designs. In that sense, calculations could have been made easier in Example 2. For instance, to estimate the average number of cars per household in the population, we could have used the same average as the one used in the sample. The 5 sampled households own a total of 6 cars, an average of 1.2 cars per household. This is the same result as that obtained using the sampling weight procedure.

Adjusting the weights

Sometimes, the sampling weights are adjusted prior to estimation. There are basically two reasons for weight adjustment:

- Adjusting for non-response: Using sampling weights for estimation works fine when you have been able to interview all selected units. In Example 2, if two of the five sampled households refused to answer or were unavailable at the time of the survey, you would only have answers for three households, thus representing only 12 of the 20 households on the street. The two non-responding units represent four households each. This means that we have no information on the number of persons or cars for 8 households on your street. In order to adjust for that, survey statisticians usually increase the weights of responding units to account for the loss of representativeness caused by non-response. The goal would be to use only the 3 units for which we have information, but still represent the 20 households on the street.

- Adjusting for external information: Sometimes, we know the actual total for one or more variables measured in the sample. In Example 3 of the Probability sampling section, a population of the 1,000 best horror movies was equally divided into 500 classic movies and 500 modern movies. Even though you knew this prior to sampling, you decided to select a simple random sample of 100 movies and ended up with 77 classic movies and 23 modern movies. Each of these movies has a weight of 10 (because you selected 1 movie out of every 10 titles). Using the answers from the survey and the sampling weight, your sample would represent a population of 770 classic movies and 230 modern movies. This could lead to inaccurate estimates. One solution would be to decrease the weight of every sampled classic and increase the weight of every sampled modern movie so that your sample gives an estimate of 500 classics and 500 modern films in the population. This should reduce the distortion caused by a 'bad' sample.

  Of course, stratifying by release date prior to sampling would have solved this problem. However, in a lot of cases, we have totals at the population level, but we don't know the attribute of each unit on the sampling frame. For example, from the Census of Population, we know how many men and women there are in a specific city, but all we have for sampling is a list of households. Thus, stratifying our population by sex would not be possible. Demographic projections by age and sex for each province are often used in social surveys to adjust sampling weights.

The weights adjusted for non-response and/or external counts are used for estimation, in the same way as the sampling weight was used in Example 1.

Other estimation methods

Using the weights to inflate the sample results is not the only estimation method that exists, but it is the simplest one and the only one that we will cover. Nevertheless, it is important to know that there exist some other methods that could lead to more precise estimates (e.g., using auxiliary information). The estimation process has to take into account the sampling design that was used. Otherwise, the resulting estimates could be severely biased.

Estimating the sampling error

As mentioned before, any estimates derived from samples are subject to what is called the sampling error. This comes from the fact that only a part of the population was observed, instead of the whole. A different sample could have come up with different results. The amount of variation that exists among the estimates from the different possible samples is what makes the sampling error. (There are roughly 14 million different combinations of 6 numbers from 1 to 49, so imagine how many ways there are to select a sample of 25,000 Canadian households!) Of course, this sampling error is unknown, since we would need to know the answer for each unit of the population in order to calculate it. Nevertheless, it can be estimated by using the survey data. The extent of the sampling error depends on many things, including the sampling method, the estimation method, the sample size and the variability of the estimated characteristic. This is why each sample estimate has its own sampling error. This error should thus be approximated for each estimated total, average, proportion, etc. produced by the survey.

Examples of estimation using an simple random sampling design

Simple random sampling is the simplest of all sampling methods. Estimation using the simple random sampling method has been studied extensively. There are simple formulas to estimate the sampling error for many statistics when simple random sampling is used, especially since it is a self-weighting design. We present here the most common estimator for a population average (mean) and total, under simple random sampling.

Estimation of the population mean

In a simple random sample, the estimate of the population mean is identical to the mean of the sample:

$$\hat{\bar{X}} = \frac{\sum x}{n}$$

where
x = an observed value

$\hat{\bar{x}}$ = estimate of the population mean
Σ
x = sum of all observed **x** values in the sample
n = number of observations in the sample.

Note: Lowercase **x** and **n** should be used if you are referring to a sample survey and upper case **X** and **N** should be used when referring to a population.

If the sample results have been summarized in a frequency table, then the estimate for the population mean is the same as the sample. Thus,

$$\hat{\bar{x}} = \frac{\Sigma x f}{\Sigma f}$$

where

x = an observed value
f = the frequency of the value (the number of times that this value have been observed in the sample)
$\hat{\bar{x}}$ = estimate of the population mean
Σ
xf = sum of all observed **xf** values (the product of the observed values times its frequency) in the sample
Σ
f = sum of the frequencies in the sample.

Example 2: A farmer randomly selects 10 eggs from a gross of 12 dozen eggs (144 eggs) he finds in his hen house. He carefully weighs each egg. The following weights were recorded in grams:

0.75, 0.70, 0.55, 0.50, 0.60, 0.65, 0.75, 0.65, 0.75, 0.50

What is the mean weight of the gross of eggs?

Using the above formula, we can determine the mean weight of all of the eggs:

$$\hat{\bar{x}} = \frac{\Sigma x}{n}$$
$$= \frac{6,4}{10}$$
$$= 0,64 \text{ grammes}$$

Estimation of the population total

For a simple random sample, the estimation formula of a total for the population is

$$\hat{x} = N \frac{\Sigma x}{n}$$

where
x = an observed value
$\hat{x}$ = estimated population total
Σ
x = sum of all observed **x** values in the sample
n = number of observations in the sample
N = total number of observations in the population.

It is just the estimate for the mean value multiplied by the number of units in the population. In the previous example, the mean weight of an egg is 0.64 grams, so it is logical to think that the total weight of the 144 eggs would be 92.16 grams (144 x 0.64 = 92.16 grams).

If sample results have been summarized in a frequency table, then the estimate formula for total population is

$$\hat{x} = N \frac{\Sigma x f}{\Sigma f}$$

where
x = an observed value
$\hat{x}$ = estimated population total
Σ
xf = sum of all observed **xf** values in the sample
Σ
f = sum of frequencies in the sample
N = total number of observations in the population.

Date Modified: 2013-07-23

Exercises

1. Do any of the following use simple random sampling? Provide a brief explanation of how each example uses this sampling method or not.
    a. bingo game
    b. Canadian election
    c. census

2. Imagine that a local clothing manufacturer has 2,700 employees. The personnel manager decides to ask the employees for suggestions on how to improve their workplace. It would take too long to survey everyone, so the manager chooses to systematically sample 300 of the employees.
    a. What would be the sampling interval?
    b. If the number 8 was your first randomly drawn number, what would be the first 5 numbers of your sample?

3. Suppose a national sports association wants to conduct a survey asking people living in Canadian capitals to choose their favourite team sport among hockey, lacrosse, curling and ringette.
    a. Copy the following table into your notebook. Fill in the missing province and capital information. (You can obtain the 2006 total population figures from Statistics Canada's Community Profiles website. In the 'Place name' space, type in the name of the capital city. Then, select a province or territory from the pull-down list. Hit the 'Search' button. When the available entries appear onscreen, choose the "city" – e.g. Regina (City). Do not use figures for census agglomerations or census metropolitan areas (CMAs).)

Table 1. Total population figures by capital cities in Canada, 2006

Capital	Territory/Province	Total population 2006 Census
	Yukon	
Yellowknife		
	Nunavut	
Victoria	British Columbia	
	Alberta	
Regina		
Winnipeg	Manitoba	
Toronto		
Québec		
	New Brunswick	
Halifax	Nova Scotia	
	Prince Edward Island	
St. John's	Newfoundland and Labrador	

    b. What is the total population of all of the Canadian capital cities combined?

    c. Table 2 features the results of the fictional survey. Review this information and answer the following questions
        i. In what cities was hockey the most popular and the least popular?
        ii. In what cities was lacrosse the most popular and the least popular?
        iii. In what cities was curling the most popular and the least popular?
        iv. In what cities was ringette the most popular and the least popular?

Table 2. Team sport preferences, selected capital cities, Canada

	Victoria	Fredericton	Regina	Yellowknife	Toronto	Québec
	%					
Hockey	10	3	58	23	71	29
Lacrosse	46	11	25	5	7	11
Curling	25	11	3	30	18	41
Ringette	19	75	14	42	4	19
Total	**100**	**100**	**100**	**100**	**100**	**100**

d. If you were to break down the data into further categories, what would you suggest?

4. 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 sex, per grade, from Kindergarten to Grade 12.

Table 3. Number of students by sex and grade level, Poplar Ridge Academy

	Kindergarten	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 Kindergarten students that should be included in the sample.

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

   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.)

Class activities

1. Choose one of the random sampling methods described in this section to survey your class on each of the following topics:
   a. average number of children in a family
   b. type of transport used to get to school
   c. number of household pets, types (i.e., cat, dog, etc.) and breed.

2. Obtain the attendance registers for two classes given by your homeroom teacher that list the name and sex of each student. Using the stratified sampling technique, survey 20% of these classes to find out the overall favourite subjects for this group of students. Use sex and class as strata characteristics.

Answers

1. The following describes the type of sampling method used in each example.
    a. A bingo game uses the simple random sampling method. All the numbers (total population) are put into a barrel and the required number (sample) are drawn at random. Each item has an equal chance of selection.
    b. A Canadian election is an example of non-random sampling (volunteer sampling) because each member of the population (18 years or older) can participate if they so desire.
    c. A census does not employ simple random sampling because every member of the target population must be included.

2. Given the sample size of 300 survey participants from a population of 2,700 employees,
    a. the sampling interval would be nine (2,700 ÷ 300 = 9).
    b. the first 5 numbers of the sample would be 8, 17, 26, 35 and 44.

3. The following is a complete table showing the population figures for Canada's provincial and territorial capitals, according to the 2006 Census of Population.
    a.

Table 1. Population of capital cities in Canada, 2006

Capital	Territory/Province	Population 2006 Census
Whitehorse	Yukon	*20,461*
Yellowknife	*Northwest Territories*	*18,700*
Iqaluit	Nunavut	*6,184*
Victoria	British Columbia	*78,057*
Edmonton	Alberta	*730,372*
Regina	*Saskatchewan*	*179,246*
Winnipeg	Manitoba	*633,451*
Toronto	Ontario	*2,503,281*
Québec	*Quebec*	*491,142*
Fredericton	New Brunswick	*50,535*
Halifax	Nova Scotia	*372,679*
Charlottetown	Prince Edward Island	*32,174*
St. John's	Newfoundland and Labrador	*100,646*

Note: The italicized entries are the missing components from Table 1 in Question 3. a).

    b. The total population size for Canadian capitals would be 5,216,928.
    c. The information from Table 1 states that:
        i. Hockey was most popular in Toronto and least popular in Fredericton.
        ii. Lacrosse was most popular in Victoria and least popular in Yellowknife.
        iii. Curling was most popular in Québec and least popular in Regina.
        iv. Ringette was most popular in Fredericton and least popular in Toronto.
    d. The data could be organized by sex and age. Respondents could also be asked to indicate whether or not they have played the sport in the past or in the present, or whether they are simply a fan. Similar questions could also be asked.

4. Given the information provided by Table 3 in Question 4:
    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 kindergarten 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. Number of students by sex and grade level needed to make a sample of 180 students, Poplar Ridge Academy

	Kindergarten	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**

Data collection

Data collectors

Individuals and organizations collect data because the information is needed. They may want information to keep records for administrative purposes, make decisions about important issues, or they may be required to pass information on to others. Whatever the specific reason, data have to be collected to provide information.

Information users

But who in society wants or needs information? Some of the many groups and organizations that use statistics include:

- **Governments:** Federal, provincial and local governments need information on the population and the economy, among other things. This information is used to develop, implement and monitor social and economic programs. 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 to measure a government's performance in making these decisions and holds it accountable if it does not meet these measurements.
- **Businesses:** Most Canadian businesses require information. This information may be about the economy of a local population or various social trends. It helps them make decisions about employing people, marketing their products and opening new offices, warehouses and factories.
- **Community groups:** These organizations need information about a wide variety of subjects, such as Aboriginal people's health and population distribution, 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.
- **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.

Statistics are often developed through a process commonly referred to as a survey. A statistical survey is developed by using well-defined concepts, methods and procedures, and compiling this information into a format that is useful such as publications or news articles. A survey involves the collection of different types of data about a particular topic of interest. The information collected can be from various units of a population (e.g., sample of television viewers or Sunday shoppers) or all units of a population (e.g., Census of Population, Census of Agriculture). It can be collected either directly from the sampled population or through the use of administrative data.

Surveys of human and non-human populations provide an important source of basic social and economical scientific knowledge. Many special interest groups (economists, sociologists, etc.) obtain grants from the government to study issues like racial violence in schools, voting behaviours, the number of children in single parent homes, etc.

Types of data collection

Data can be collected using three main types of surveys: censuses, sample surveys, and administrative data. Each has advantages and disadvantages. As students, you may be required to collect data at some time. The method you choose will depend on a number of factors.

Census

A 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 class census. There are various reasons why a census may or may not be chosen as the method of data collection:

Advantages (+)

Sampling variance is zero: There is no sampling variability attributed to the statistic because it is calculated using data from the entire population.

Detail: Detailed information about small sub-groups of the population can be made available.

Disadvantages (−)

Cost: In terms of money, conducting a census for a large population can be very expensive.

Time: A census generally takes longer to conduct than a sample survey.

Response burden: Information needs to be received from every member of the target population.

Control: A census of a large population is such a huge undertaking that it makes it difficult to keep every single operation under the same level of scrutiny and control.

Sample survey

In a sample survey, only part of the total population is approached for data. 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. Reasons one may or may not choose to use a sample survey include:

Advantages (+)

Cost: A sample survey costs less than a census because data are collected from only part of a group.

Time: Results are obtained far more quickly for a sample survey, than for a census. Fewer units are contacted and less data needs to be processed.

Response burden: Fewer people have to respond in the sample.

Control: The smaller scale of this operation allows for better monitoring and quality control.

Disadvantages (−)

Sampling variance is non-zero: The data may not be as precise because the data came from a sample of a population, instead of the total population.

Detail: The sample may not be large enough to produce information about small population sub-groups or small geographical areas.

Administrative data

Administrative data are collected as a result of an organization's day-to-day operations. Examples include data on births, deaths, marriages, divorces and car registrations. For example, prior to being issued a marriage license, a couple must provide the registrar with information about their age, sex, birthplace, address and previous marital status. These administrative files can be used later as a substitute for a sample survey or a census.

Advantages (+)

Sampling variance is zero: There is no variability attributed to the statistic because it was calculated using data from the entire population.

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

Simplicity: Administrative data may eliminate the need to design a census or survey and the associated work.

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

Disadvantages (−)

Flexibility: Data items may be limited to essential administrative information, unlike a survey.

Population: Data are limited to the population on whom the administrative records are kept.

Change over time: Definitions are created to serve specific purposes, but often change and evolve over time. The statistician must understand that there is a possibility of change to the definitions of these files.

Concepts and definitions: The definitions are established by those who create and manage the file for their own purposes. For example, income definitions may not include everything a user expects to see.

Data quality: The emphasis placed on data quality may differ from organization to organization. This may be evident when someone relies on data collected from another organization.

Data collection methods

Personal interviews

Face-to-face: involves trained interviewers visiting people to collect questionnaire data. It is a good approach for ensuring a high response rate to a sample survey or census, and trained interviewers gather better quality data. However, there are some disadvantages to this approach. Respondents may not always be available for interviews and the travel costs of the interviewer could be high.

Computer Assisted Personal Interviewing (CAPI): is a form of personal interviewing, but instead of completing a questionnaire, the interviewer brings along a laptop or hand-held computer to enter the information directly into the database. This method saves time involved in processing the data, as well as saving the interviewer from carrying around hundreds of questionnaires. However, this type of data collection method can be expensive to set up and requires that interviewers have computer and typing skills.

Telephone interviews

Telephone: involves trained interviewers phoning people to collect questionnaire data. This method is quicker and less expensive than face-to-face interviewing. However, only people with telephones can be interviewed (about 98% of the Canadian population), and the respondent can end the interview very easily!

Computer Assisted Telephone Interviewing (CATI): is a type of telephone interview, but with the interviewer keying respondent answers directly into a computer. This saves time involved in processing data, but can be expensive to set up, and requires interviewers to have computer and typing skills. Statistics Canada uses this approach for many of its surveys such as the Youth In Transition Survey, the Monthly Survey of Manufacturing, the General Social Survey and the Workplace Employee Survey.

Self-completed

Mail survey: a common method of conducting Statistics Canada's economic surveys. It is a relatively inexpensive method of collecting data, and one that can distribute large numbers of questionnaires in a short time. It provides the opportunity to contact hard-to-reach people, and respondents are able to complete the questionnaire in their own time. Mail surveys do require an up-to-date list of names and addresses, however. In addition, there is also the need to keep the questionnaire simple and straightforward.

A major disadvantage of a mail survey is that it usually has lower response rates than other data collection methods. This may lead to problems with data quality. Also, people with a limited ability to read or write English or French may experience problems.

Hand-delivered questionnaire: a self-enumerated survey where questionnaires are hand-delivered to people and mailed back by the respondent after completion. This method usually results in better response rates than a mail survey, and is particularly suitable when information is needed from several household members. The hand-delivered with pickup method has been used by Statistics Canada's Census of Population. The hand-delivered with respondent mail-back method can reduce the cost of collecting forms and gives a greater sense of privacy for respondents concerned with someone entering their home or business to collect the forms.

Other methods

Electronic Data Reporting (EDR): Electronic forms have been available at Statistics Canada for some surveys (mainly for business surveys) since the early 1990s. Although this type of data reporting is still quite rare, it gives the respondents the option of choosing how they would like to report the data: filling out the usual paper questionnaire or using the electronic version. Because the technology evolves so quickly, remaining up-to-date with good and secure applications requires major investments. Statistics Canada keeps up its efforts in this area.

The Internet: The growing popularity of the Internet brought a major shift in Electronic Data Reporting (EDR). It is hard to find a quick and easy way of reporting answers through the Internet without sacrificing any of Statistics Canada's principles concerning confidentiality, privacy and data quality. The Agency has begun introducing pilot projects for a diverse range of important surveys involving respondents from households, universities, businesses, and federal departments. Pilot projects include, most recently, the 2001 Census of Population, the Annual Retail and Wholesale Trade Survey, the Unified Enterprise Survey and the Business Payroll Survey.

Other methods include direct observation, such as that used in pricing surveys, or the use of existing administrative records. The choice of method depends on various factors: complexity and length of questionnaire, sensitivity of requested information, geographical dispersion of survey population, cost and time frame.

Often the most satisfactory collection strategy uses a combination of methods. For example, mail surveys have proven to be quite efficient when designed as a follow-up for those who did not respond by telephone interview.

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-barreled?
- Are any questions leading or loaded?
- Should the questions be open- or close-ended? If the questions are close-ended are the response categories mutually exclusive and 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, the status of the survey (voluntary or mandatory) and any existing 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 completely 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. Testing helps discover poor wording or ordering of questions; identify errors in the questionnaire layout and instructions; determine problems caused by the respondent's inability or unwillingness to answer the questions; suggest additional response categories that can be pre-coded on the questionnaire; and 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.

Role of interviewers

It is important to note that not all persons who collect data are interviewers. In some instances, the data are collected by having people go into grocery stores or clothing stores on a monthly basis. They record the price of a given list of goods and services on hand-held devices and then they report their data back to Statistics Canada.

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, when someone is employed to collect data they may need:

- good communication skills;
- a confident and professional appearance; and
- use of a car and telephone.

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. (See Questionnaire design section.)

In addition, it is important that the interviewer have 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 for 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.

Exercises

Data collection exercise

1. The first census to take place after Confederation was held in 1871. Which of the four original Canadian provinces were enumerated?
    a. Nova Scotia, Quebec, Ontario and Manitoba
    b. New Brunswick, Quebec, Manitoba and Newfoundland
    c. Nova Scotia, New Brunswick, Quebec and Ontario
    d. None of the above

2. When collecting data, why is it sometimes better to conduct a sample survey than a census?

3. Suggest reasons why data would be collected on the following topics:
    a. burglaries
    b. causes of death
    c. climate
    d. forests
    e. immigration
    f. schools

4. List some of the things you would need to consider when choosing a data collection method.

5. Given some of your answers to Question 4, decide as a class which method of data collection you would employ to gather data on the following topics:
    a. the music tastes of your class
    b. the average height of your class
    c. the time your parents spend each week doing housework
    d. the attitude of Canadian students toward the environment

6. Are there some topics for which data should not be collected? For example, data on people's health or political beliefs? Discuss as a class where you would 'draw the line' in deciding if issues are too sensitive to ask about.

    What factors would help researchers decide? For example, is the information to be collected of national importance?

Questionnaire exercise

The following questionnaire has been designed to collect information about the users of Greenwood's Public Library. There are some problems with this questionnaire. Read through and see how many you can explain.

Name:

Telephone: Introduction

1. How often do you use the services offered by the library?

2. How many books or periodicals have you borrowed from the library?

- 0
- 1-5
- 5-10
- 10-15
- 20-50
- 50-10

3. The last time you used the library, what was the purpose of your visit?

- search for a book

- search for a periodical
- get information from a librarian
- study peacefully

4. Were your needs satisfied?

- Yes
- No

5. How satisfied are you with the quality of service provided by the library and the attitude of the library staff?

1 2 3 4 5

6. What do you dislike about the library?

7. Are there any improvements which could be made to the library in order to provide better service?

- Yes
- No

8. Do you approve or disapprove of the recent proposals made by the Library Management Review Committee, such as the proposal to double fines for overdue books?

- Approve
- Disapprove

(Go to **Question 11**)

9. Are you aware of these proposals?

- Yes
- No

10. Why do you disapprove of these proposals?

11. Are you against not having longer opening hours?

- Against
- Not against

12. Level of education

Thank you for your help!

Answers

Data collection answers

1. The answer is c) Nova Scotia, New Brunswick, Quebec and Ontario.

2. It is sometimes better to take a sample survey than a census because it is less expensive, quicker to undertake, serves specialized needs, lessens respondent burden and may only require information from a certain segment of the population.

3.

4. When choosing a data collection method, you should look at the following elements:
    - cost (budget);
    - time;
    - size of population; and
    - personnel required to perform the chosen method.

Discuss the other questions with your class.

Questionnaire answers

Outlined below are just some of the possible problems with the Greenwood Public Library questionnaire. Did you find other problems with this questionnaire? Discuss as a class.

Introduction

There is no introduction telling the respondent what the purpose of the survey is and what will be done with the results. This would help readers to "screen themselves out" and not fill in the survey if they are not library users.

If you are looking for open and honest results, you should not ask for name and telephone number. People like to know that the information they provide is confidential and will be kept that way.

Question 1:

This question is *open* and may be better worded as a *closed* question in order to be able to compare people's responses. Also, this question should address the reference period being considered (i.e., how often did one visit in the last 12 months, or the last month, or the last week).

Question 2:

The question asks about borrowing books or periodicals. What if you borrowed both? What response would you select if you had borrowed five books and five periodicals? Would you select "1–5" or "5–10"? The other issue with this question is the reference period. Has the borrowing taken place over the last year, the last 10 years or the last month?

Question 3:

This closed question only gives four answers to choose from. However, there may be other reasons for going to the library. You could add an "Other" category with a "please specify" space, or a longer list of reasons including "social activity" or "book signing". As well, you might add an instruction to this question that states "mark all that apply". This way the respondent can choose more than one reason.

Question 4:

The question is fine but does not go far enough in getting the details. The library likely needs to know why respondents' needs were not satisfied. This question should include a second part that asks respondents to explain how their needs were not satisfied.

Question 5:

This is a double-barreled question. The first problem is that you are asking the respondent two different things in one question. The second problem is that there are no instructions with the response categories provided. Do respondents rank their degree of satisfaction? If so, what does 1 represent? Does 5 stand for poor or excellent?

Question 6:

This is an open question, which makes the various possible responses difficult to code and tabulate. A list of answers (including a final "Other" choice with a space to specify the response) would make this question a lot easier to answer and the results easier to tabulate.

Question 7:

This question asks for a "Yes" or "No" response but the respondent may not know, or may have no opinion regarding any improvements to the library.

Questions 8, 9 and 10:

The first problem with these questions is that Question 9 ("Are you aware of these proposals?") should come before Question 8 ("Do you approve or disapprove of the recent proposals...?") since Question 9 is only relevant to Question 8 if the answer is "Yes".

In Question 8, the respondents are asked whether they approve or disapprove with the entire set of proposals. What if the respondent agreed with some of the proposals and disagreed with others? The available responses do not give a true measure of the proposals. The wording of Question 8 should be changed in order to avoid bias.

The "Go to" part of Question 8 also contains a problem. It sends the respondent to Question 11, therefore respondents who answered "disapprove" to Question 8 would not be given the opportunity to respond to Questions 9 or 10. The instruction should read "(Go to Question 10)".

Question 11:

This question presents the reader with a double negative. Because it is not entirely clear what the question is asking, some respondents could interpret it in different ways. You could reword it to say: "Are you in favour of the library extending the hours of operation?" Include a list of responses such as "__Yes __No __ Don’t care".

Question 12:

This particular question is sensitive and respondents may not feel inclined to answer truthfully or at all. It may be best not to ask this question. If, however, the level of education is relevant and necessary information, then a closed question would allow better quantitative analysis.

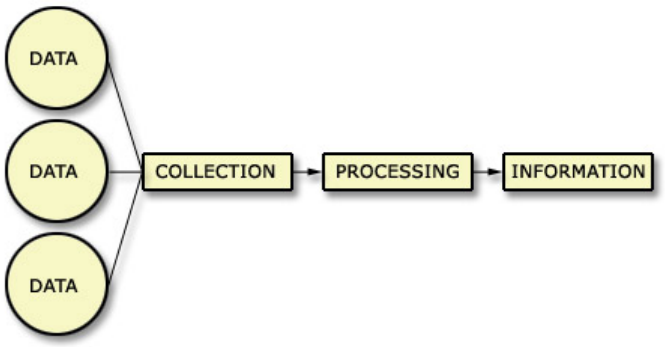
Data processing

Data are facts or figures from which conclusions can be drawn. When data have been recorded, classified, and organized, related or interpreted within a framework so that meaning emerges, they become information. There are several steps involved in turning data into information, and these steps are known as *data processing*. This chapter looks at data processing and how computers perform these steps quickly and easily.

Introduction

The simplified flowchart below shows how raw data are transformed into information. Data processing takes place once all of the relevant data have been collected. They are gathered from various sources and entered into a computer where they can be processed to produce information (output).

Figure 1. Data processing flowchart



Data processing includes the following steps:

- Coding
- Capture
- Editing
- Imputation
- Quality control
- Producing results

Coding

First, before raw data can be entered into a computer, they must be coded. In order to do this, survey responses must be labeled, usually with simple, numerical codes. This can be done by the interviewer in the field or even by an office employee. The data coding step is important because it makes data entry and data processing easier.

Surveys have two types of questions—*closed questions* and *open questions*. The responses to these questions affect the type of coding performed. A closed question means that only a fixed number of predetermined survey responses are allowed. These responses will have already been coded.

The following question asked in the 1998 Time Use Survey (Sport), 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

An open question implies that any response is allowed, making subsequent coding more difficult. In order to code an open question, the processor must sample a number of responses, and then design a code structure that includes all possible answers.

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

What sports do you participate in?

Specify (28 characters)_____

In the Census and almost all other surveys, the codes for each question field are pre-marked on the questionnaire. When it comes time to process the questionnaire, the codes are entered directly into the database and are prepared for data capturing. The following is an example of pre-marked coding:

What language does this person speak most often at home?

<18/> English

<19/> French

<20/> Other — Specify _____

Statistique Canada a élaboré des codes uniformes, désignés classifications ou normes, pour aider à répartir les personnes, les endroits et les choses en groupes spécialisés. Se reporter à la section sur la classification pour obtenir plus de renseignements à ce sujet.

Automated coding systems

Statistics Canada is constantly testing programs that will automate repetitive and routine tasks. Thus, coding is a prime candidate for this type of automation.

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. 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 Standard Industrial Classification System and the Standard Occupation Classification System. The rejected records (those that do not have a match with the written response) are the only data to be manually coded.

The 1991 Census of Population used an automated coding system to code language, religion and ethnic origin information. These codes are not part of a standard classification system, but they are a standard used by the Census to facilitate analysis and tabulations.

The next step in data processing is inputting the coded data into a computer database. This method is known as *data capture*.

Data capture

This is the process by which data are transferred from a paper copy (questionnaires and survey responses) to an electronic file. The responses are then put into a computer.

Before this procedure takes place, the questionnaires must be *groomed* (prepared) for data capture. In this processing step, the questionnaire is reviewed by someone to ensure that all of the minimum required data have been reported, and that they are decipherable. This grooming is usually performed during extensive automated edits.

There are several methods used for capturing data:

- **tally charts** are used to record data such as the number of occurrences of a particular event and to develop frequency distribution tables. Tally charts are used only with small amounts of data. This form of data capture would never be used for a census. However, in the instance where a hand count is required (e.g., election ballots), a tally chart could be considered quite useful.
- **batch keying** is one of the oldest methods of data capture. It uses a computer keyboard to type in the data. This process is very practical for high-volume entry where fast production is a requirement. No editing procedures are necessary but there must be a high degree of confidence in the editing program. Also, validity and range edits need to be implemented to ensure quality keying. This does not mean the data are being re-edited, but if a field is numeric and alpha characters are entered instead, the error will be flagged. This approach can be beneficial when used for large surveys with many questions and edits.
- **interactive capture** is often referred to as *intelligent keying*. Usually, captured data are edited before they are imputed. However, this method combines data capture and data editing in one function. Although interactive capture is slower, it is a very effective approach to use when there is a lot of interdependency between questions. This process requires knowledge of editing procedures, as the errors need to be corrected right away. Interactive capture also reduces the number of documents handled, as the edits are made directly on a computer.
- The edits performed during an interactive capture can be preliminary data checks or basic edits (such as total, subtotal and value checks), or they can consist of full edits applied to the entire questionnaire. (One of these systems is referred to as the computer-assisted telephone interviewing process discussed in the chapter on Data collection).
- **optical character readers** or bar-code scanners, are able to recognize alpha or numeric characters. These readers scan lines and translate them into the program.

  These bar-code scanners are quite common and often seen in department stores. They can take the shape of a gun or a wand. The gun scanner is simple to program and can verify the validity of what has been scanned. The Census of Agriculture uses this scanning method for its document-control system. A printer prints out bar codes to label the questionnaires. Then the bar coder scans the label and translates the lines into numbers that are entered into the document-control file.

  The wand or pen type of scanner reads identification numbers. This scanning method requires a slow and steady hand in order to ensure that the wand is aligned properly with the printed characters.
- **magnetic recordings** allow for both reading and writing capabilities. This method may be used in areas where data security is important. The largest application for this type of data capture is the PIN number found on automatic bank cards.

A computer keyboard is one of the best known input (or data entry) devices in current use. In the past, people performed data entry using punch cards or paper tape.

Some modern examples of data input devices are

- optical mark reader
- bar-code reader
- scanner used in desktop publishing
- light pen
- trackball
- mouse

Did you know that ...

- A new objective for 2001 was to create an image retrieval system giving access to the images (pictures) of all of the census questionnaires and visitation records, so that subsequent processes requiring access to original census forms would not have to handle the thousands of boxes and paper documents, as in previous censuses.
- Users will have access to more information free of charge on the internet through Statistics Canada's Website.
- In 2001, for the first time, the census collects information on same-sex couples, as well as information on language of work.
- The respondent's guide that previously accompanied the short questionnaire was no longer printed, saving some 215,460 kg of paper.
- Approximately 13.2 million questionnaires were keyed using 5 billion key strokes.
- Alternative means of collection and processing included the Internet, by which Canadians could file their questionnaires on-line; automated editing processes; and a computer-assisted telephone interview application to name several.

Once data have been entered into a computer database, the next step is ensuring that all of the responses are accurate. This method is known as *data editing*.

Data editing

Data should be edited before being presented as information. This action ensures that the information provided is accurate, complete and consistent. No matter what type of data you are working with, certain edits are performed on all surveys. Data editing can be performed manually, with the assistance of computer programming, or a combination of both techniques. It depends 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 compatibility of data.

We might ask the question "Why are there 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.
- An interviewer could have miscoded or misunderstood a written response.
- An interviewer could have forgotten to ask a question or record the answer.
- A respondent could have provided inaccurate responses.

Always keep in mind the objectives of data editing:

- to ensure the accuracy of data;
- to establish the consistency of data;
- to determine whether or not the data are complete;
- to ensure the coherence of aggregated data; and
- 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.

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 is 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 surveys relating to the same sort of variables or characteristics are used in order to establish some of the rules for editing data.

Data editing types

There are several types of data edits available: They 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 reporting time is within the specified limits.
- **Range edits** are similar to validity edits in that they look at one field at a time. The purpose of this type of edit is to ensure that the values, ratios and calculations fall within the pre-established limits.
- **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 item 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 into 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 and/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 errors

Data editing should detect and minimize errors, such as:

- unasked questions;
- unrecorded answers;
- inappropriate responses.

An inaccurate response can occur as a result of carelessness or a deliberate effort to give misleading answers. It can also occur if some of the answers require mathematical calculations. For example, converting days into hours or annual income into weekly income increases the possibility of making mistakes.

Example 1 – Inaccurate responses

This example of data editing shows how an inaccurate response can occur. Carefully read the following questions and answers based on the questions asked in Statistics Canada's Labour Force Survey form. Can you detect the error in the respondent's answers?

Question 151 - Excluding overtime, how many **paid** hours does *Person 1* work per week?
Answer - 40

Question 153 - Last week, how many hours was *Person 1* away from this job because of vacation, illness, or any other reason?
Answer - 0

Question 155 - Last week, how many hours of paid overtime did *Person 1* work at this job?
Answer - 4

Question 156 - Last week, how many extra hours without pay did *Person 1* work at this job?
Answer - 0

Question 157 - Last week, how many hours did *Person 1* actually work at the main job?
Answer - 40

Question 151 shows that Person 1 normally works 40 hours per week. Question 153 shows the respondent had no time off the previous week, and Question 155 shows that, in fact, some overtime was worked. However, Question 157 gives us the answer that all of this amounted to 40 hours worked for the week! The actual response to Question 157 should be 44 hours.

The answers to individual questions look acceptable. It is only by comparing them with each other that we find one or more of the answers to be wrong.

Cross-referencing, a form of a consistency edit, is only one type of data editing that compares the answers of various questions. Cross-referencing can be performed manually or with the use of editing software.

This edit indicates that further action should be taken to ensure an accurate response in the above example; the interviewer will need to get in touch with the household and verify the number of hours worked by Person 1.

In computer-assisted personal or telephone interviews, the interviewer would receive an electronic warning when trying to enter 40 as a response to Question 157. The interviewer could then immediately double-check the answer with the respondent. This system is much faster, and eliminates the burden of trying to contact the respondent later on.

Editing as a management tool

The editing process can also be a valuable tool in assessing the quality of the data by indicating the required modifications. By indicating potential causes of problems, editing can also be an effective way of avoiding the need to repeat the survey.

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.

Non-response and invalid data definitely impact the quality of the survey results. Imputation resolves the problems of missing, invalid or incomplete responses identified during editing, as well as any editing errors that might have occurred. At this stage, all of the data are screened for errors because respondents are not the only ones capable of making mistakes; errors can also occur during coding and editing.

Imputation procedures are designed to fill in the gaps. So, 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.

The imputation procedures are decided upon during the planning and development stages of a survey. Some problems are eliminated earlier through contact with the respondent or by manually studying the questionnaire, but it is generally impossible to resolve all problems due to concerns of response burden, cost and timeliness. Thus, the imputation procedure is used to handle the remaining edit failures.

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.

There are several approaches to consider when imputing data. Usually, deductive imputation is 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).

Some other types of imputation methods include:

- **hot deck** uses other records as 'donors' in order 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.
- **substitution** relies on the availability of comparable data. Imputed data can be extracted from the respondent's record from a previous cycle of the survey, or the imputed data can be taken from the respondent's alternative source file (e.g. administrative files or other survey files for the same respondent). This is often difficult to do because, in many cases, there is no other information available than the information provided in the current survey.
- **estimator** uses information from other questions or from other answers (from the current cycle or a previous cycle), and through mathematical operations, derives a plausible value for the missing or incorrect field.

  The simplest of the estimator methods is the *mean imputation*. With this approach, a missing field is filled with the average value 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 recorded average income in that individual's province for the same occupation with the same level of experience as the respondent. There are other, more sophisticated estimator methods available.
- **cold deck** makes use of a fixed set of values, which covers all of the data items. These values can be constructed with the use of historical data, subject-matter expertise, etc. A 'perfect' questionnaire is created in order to answer complete or partial imputation requirements.
- The donor can also be found through a method called **nearest neighbour imputation**. In this case, some sort of criteria must be developed to determine which responding unit is 'most like' 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.

The method of imputation can vary from survey to survey and, depending on unique or particular circumstances, sometimes even within the same survey. These methods can be applied either manually or with the use of an automated system. The imputed value is determined by calling the respondent or is based on the judgment of a subject-matter specialist. To help facilitate this, Statistics Canada has written specialized programs to impute data based on the methodological input of experienced statisticians who have analysed the survey and suggested approaches on how best to impute meaningful data.

Imputation methods can be performed automatically, manually or in combination. Done properly, imputation limits the biases caused by not having a complete and accurate record; contains an audit trail for evaluation purposes; and ensures that the imputed records are internally consistent. A good imputation procedure is automated, objective and efficient.

Data quality

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
- act on the differences.

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.

Outlined below are some of the key 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
- uses 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

Quality management in statistical agencies

The quality of the data must be defined and assured in the context of being 'fit for use'. Whether or not data and statistical information are fit for use will depend on the intended function of the data and the fundamental characteristics of quality. It also depends on the users' expectations of what they consider to be useful information.

There is no standard definition among statistical agencies for the term *official statistics*. There is a generally accepted, but evolving, range of quality issues underlying the concept of 'fitness for use'. These elements of quality need to be considered and balanced in the design and implementation of an agency's statistical program.

So, how does Statistics Canada define quality? The following is a list of the elements of quality:

- **relevance:** The relevance of statistical information reflects the degree to which it meets the real needs of clients. It is concerned with whether the available information sheds light on the issues that are important to users. Assessing relevance is subjective and depends upon the varying needs of users. The Agency's challenge is to weigh and balance the conflicting needs of current and potential users to produce a program that goes as far as possible in satisfying the most important needs within given resource constraints.

- **accuracy:** The accuracy of statistical information is the degree to which the information correctly describes the phenomena it was designed to measure. It is usually characterized in terms of error in statistical estimates and is traditionally decomposed into bias (systematic error) and variance (random error) components. It may also be described in terms of the major sources of error that potentially cause inaccuracy (e.g., coverage, sampling, nonresponse, response).

- **timeliness:** The timeliness of statistical information refers to the delay between the reference point (or the end of the reference period) to which the information pertains, and the date on which the information becomes available. It is typically involved in a trade-off against accuracy. The timeliness of information will influence its relevance.

- **accessibility:** The accessibility of statistical information refers to the ease with which it can be obtained from the Agency. This includes the ease with which the existence of information can be ascertained, as well as the suitability of the form or medium through which the information can be accessed. The cost of the information may also be an aspect of accessibility for some users.

- **interpretability:** The interpretability of statistical information reflects the availability of the supplementary information and metadata necessary to interpret and utilize it appropriately. This information normally includes the underlying concepts, variables and classifications used, the methodology of data collection and processing, and indications or measures of the accuracy of the statistical information.

- **coherence:** The coherence of statistical information reflects the degree to which it can be successfully brought together with other statistical information within a broad analytic framework and over time. The use of standard concepts, classifications and target populations promotes coherence, as does the use of common methodology across surveys. Coherence does not neccessarily imply full numerical consistency.

These elements of quality tend to overlap, often in a confounding manner. Just as there is no single measure of accuracy, there is no effective statistical model for bringing together all these characteristics of quality into a single indicator. Also, except in simple or one-dimensional cases, there is no general statistical model for determining whether one particular set of quality characteristics provides higher overall quality than another.

Achieving an acceptable level of quality is the result of addressing, managing and balancing over time the various factors or elements that constitute better quality. Paying attention to the program objectives, the major uses of the data, costs, and conditions and circumstances that affect quality and user expectations is also important in determining an acceptable level of quality. Since the elements of quality have a complex relationship, an action taken to address or modify one aspect of quality tends to affect the other elements. Thus, the balance between these factors may be altered in ways that cannot readily be modeled or adequately quantified in advance. The decision and actions that achieve this balance are based on knowledge, experience, reviews, feedback, consultation and, inevitably, judgment.

Producing results

After editing, data may be processed further to produce a desired output. The computer software used to process the data will depend on the form of output required.

Software applications for word processing, desktop publishing, graphics (including graphing and drawing), programming, databases and spreadsheets are commonly used. The following are some examples of ways that software can produce data:

- **Spreadsheets** are programs that automatically add columns and rows of figures, calculate means, and perform statistical analyses. They can be used to create financial worksheets (such as budgets or expenditure forecasts), balance accounts, and analyse costs. Charts and graphs can be created to show the significance of a selection of data. They can be displayed in a number of ways: bar graphs, line graphs, and circle graphs/pie charts are just a few examples of the visual data that can be produced.
- **Databases** are electronic filing cabinets. They systematically store data for easy access, and produce summaries, aggregates or reports. A database program should be able to store, retrieve, sort and analyse data.
- **Specialized programs** can be developed to edit, clean, impute and process the final tabular output. This method offers 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 then produce publishable final results.

Computer output may be used in a variety of ways. It can be stored for future retrieval and use. It can be laser-printed onto paper as tables or charts, or even put onto transparent slides for overhead projector use. The output can also be saved onto electronic medium for use in portable and desktop computers. Lastly, data can be sent to others as an electronic file via the Internet.

Output is usually governed by the need to communicate specific information to a specific audience. The only limit to the different forms of output you can produce is the different types of output devices currently available. 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 will the audience best understand it?

Exercise

1. From what you already know about the statistical process, place the following steps in the correct order:
   - processing
   - collection
   - information
   - data

2. List all of the steps involved in data processing and write a brief description of each.

3. Find out what kinds of data input devices are present at your school. Do these devices require special skills to operate? Discuss as a class.

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

5. The following are some examples of incorrect survey responses. Find the errors in each example and explain them.

   a.

Question 5a

Question	Response
(i) **How many rooms are there in this dwelling?** **Include** kitchen, bedrooms, finished rooms in attic or basement, etc. **Do not count** bathrooms, halls, vestibules and rooms used solely for business purposes.	11
(ii) **How many of these rooms are bedrooms?**	Yes

   b.

Question 5b

Question	Response
What is your present marital status? *Mark **one** circle only.*	• ( ) Legally married (and not separated) • ( ) Separated, but still legally married • (x) Divorced • ( ) Widowed • (x) Never married (single)

   c.

Question 5c

Question	Response
How did you get to work on Census Day 2001? *If you used more than one method of transportation, mark all relevant circles.*	• ( ) Car, truck or van—as a driver • ( ) Car, truck or van—as a passenger • ( ) Public transit (*e.g., bus, street car, subway, light rail transit, commuter train, ferry*) • ( ) Walked to work • ( ) Bicycle • (x) Motorcycle • ( ) Taxicab • ( ) Other method • (x) Worked from home • (x) Did not go to work

6. Answer the following questions:
    - Does your school use coding procedures to produce data?
    - Are the coding procedures performed by a person or by a machine?
    - If the coding procedures are manual, would it be better to automate them? Why?

Answers

1. The correct order for these items is
    - data
    - collection
    - processing
    - information

2. The data processing steps include coding responses, capturing data, editing data, imputing data, ensuring quality and producing results.

3.

4. 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.

5.
    a. The error in this survey questionnaire is found in the respondent's answer. The respondent should have entered the number of bedrooms in part (ii) instead of the response "yes".
    b. There are two errors in this example. First, the question asked the respondent to mark only *one* of the provided answers. Instead, the respondent marked two answers. The second error is that both the "divorced" and "never married" responses were marked. No one can be divorced if they haven't married!
    c. In this example, the respondent claims that he or she went to work on a motorcycle. But the respondent also checked off the "did not go to work" circle. If the person did not go to work, then he or she could not have traveled to work by motorcycle.

    However, the respondent could legitimately mark both the "motorcycle" and "work from home" fields, because the person could have been driving to his or her home (a possible work place) from somewhere else.

Organizing data

After being collected and processed, data need to be organized to produce useful information. When organizing data, it helps to be familiar with some of the definitions. This chapter outlines those definitions and provides some simple techniques for organizing and presenting data.

Variables

The word **variable** is often used in the study of statistics, so it is important to understand its meaning. A variable is a characteristic that may assume more than one set of values to which a numerical measure can be assigned.

Height, age, amount of income, province or country of birth, grades obtained at school and type of housing are all examples of variables. Variables may be classified into various categories, some of which are outlined in this section.

Categorical variables

A categorical variable (also called qualitative variable) is one for which each response can be put into a specific category. These categories must be mutually exclusive and exhaustive. Mutually exclusive means that each possible survey response should belong to only one category, whereas, exhaustive requires that the categories should cover the entire set of possibilities. Categorical variables can be either nominal or ordinal.

Nominal variables

A **nominal variable** is one that describes a name or category. Contrary to ordinal variables, there is no 'natural ordering' of the set of possible names or categories. Sex and type of dwelling are examples of nominal variables. In Table 1, the variable "mode of transportation for travel to work" is nominal because it describes the category of transportation.

Table 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

Source: 2001 Census: analysis series catalogue no. 96F0030XIE2001010

Ordinal variables

An **ordinal variable** is a categorical variable for which the possible categories can be placed in a specific order or in some 'natural' way. In Table 2, the variable 'behaviour' is ordinal because the category 'Excellent' is better than the category 'Very 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 2. Student behaviour ranking

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

Numeric variables

A **numeric variable**, also known as a quantitative variable, is one that can assume a number of real values—such as age or number of people in a household. However, not all variables described by numbers are considered numeric. For example, when you are asked to assign a value from 1 to 5 to express your level of satisfaction, you use numbers, but the variable (satisfaction) is really an ordinal variable.

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. Examples of a **continuous variable** are distance, age and temperature.

The measurement of a continuous variable is restricted by the methods used, or by the accuracy of the measuring instruments. For example, the height of a student is a continuous variable because a student may be 1.6321748755... metres tall.

However, when the height of a person is measured, it is usually measured to the nearest centimetre. Thus, this student's height would be recorded as 1.63 m.

Note: To make them easier to handle, continuous variables are usually grouped into "class intervals", which will be discussed later in this chapter. Grouping variables is part of the process of organizing data so that they become useful information.

Discrete variables

As opposed to a continuous variable, a **discrete variable** can only take a finite number of real values. 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).

Discrete variables may also be grouped. Again, grouping variables makes them easier to handle.

Note: Measurement of a continuous variable is always a discrete approximation.

Frequency distribution tables

The *frequency* (**f**) of a particular observation is the number of times the observation occurs in the data. The *distribution* of a variable is the pattern of frequencies of the observation. Frequency distributions are portrayed as frequency tables, histograms, or polygons.

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.

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 1. Frequency table for the number of cars registered in each household

Number of cars (x)	Tally	Frequency (f)
0	IIII	4
1	𝍸 I	6
2	𝍸	5
3	III	3
4	II	2

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 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

For more information on how to make cumulative frequency tables, see the section on Cumulative frequency and Cumulative percentage.

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.

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 3. Life of AA batteries, in minutes

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

Relative frequency and percentage frequency

An analyst studying these data 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

Example 4 – Constructing relative frequency and percentage frequency tables

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. 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	0.10	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	0.10	10
430–439	1	0.03	3
Total	**30**	**1.00**	**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.

Keep in mind that these analytical statements have assumed that a representative sample was drawn. In the real world, an analyst would also refer to an estimate of variability (see section titled Measures of spread) to complete the analysis. For our purpose, however, it is enough to know that frequency distribution tables can provide important information about the population from which a sample was drawn.

Stem and leaf plots

A stem and leaf plot, or stem plot, is a technique used to classify either *discrete* or *continuous* variables. A stem and leaf plot is used to organize data as they are collected.

A stem and leaf plot looks something like a bar graph. Each number in the data is broken down into a stem and a leaf, thus the name. The stem of the number includes all but the last digit. The leaf of the number will always be a single digit.

Elements of a good stem and leaf plot

A good stem and leaf plot

- shows the first digits of the number (thousands, hundreds or tens) as the *stem* and shows the last digit (ones) as the *leaf*.
- usually uses whole numbers. Anything that has a decimal point is rounded to the nearest whole number. For example, test results, speeds, heights, weights, etc.
- looks like a bar graph when it is turned on its side.
- shows how the data are spread—that is, highest number, lowest number, most common number and outliers (a number that lies outside the main group of numbers).

Tips on how to draw a stem and leaf plot

Once you have decided that a stem and leaf plot is the best way to show your data, draw it as follows:

- On the left hand side of the page, write down the thousands, hundreds or tens (all digits but the last one). These will be your stems.
- Draw a line to the right of these stems.
- On the other side of the line, write down the ones (the last digit of a number). These will be your leaves.

For example, if the observed value is 25, then the stem is 2 and the leaf is the 5. If the observed value is 369, then the stem is 36 and the leaf is 9. Where observations are accurate to one or more decimal places, such as 23.7, the stem is 23 and the leaf is 7. If the range of values is too great, the number 23.7 can be rounded up to 24 to limit the number of stems.

In stem and leaf plots, tally marks are not required because the actual data are used.

Not quite getting it? Try some exercises.

Example 1 – Making a stem and leaf plot

Each morning, a teacher quizzed his class with 20 geography questions. The class marked them together and everyone kept a record of their personal scores. As the year passed, each student tried to improve his or her quiz marks. Every day, Elliot recorded his quiz marks on a stem and leaf plot. This is what his marks looked like plotted out:

Table 1. Elliot's scores on the basic facts quiz last year

Stem	Leaf
0	3 6 5
1	0 1 4 3 5 6 5 6 8 9 7 9
2	0 0 0 0

Analyse Elliot's stem and leaf plot. What is his most common score on the geography quizzes? What is his highest score? His lowest score? Rotate the stem and leaf plot onto its side so that it looks like a bar graph. Are most of Elliot's scores in the 10s, 20s or under 10? It is difficult to know from the plot whether Elliot has improved or not because we do not know the order of those scores.

Try making your own stem and leaf plot. Use the marks from something like all of your exam results last year or the points your sports team accumulated this season.

The main advantage of a stem and leaf plot

The main advantage of a stem and leaf plot is that the data are grouped and all the original data are shown, too. In Example 3 on battery life in the Frequency distribution tables section, the table shows that two observations occurred in the interval from 360 to 369 minutes. However, the table does not tell you what those actual observations are. A stem and leaf plot would show that information. Without a stem and leaf plot, the two values (363 and 369) can only be found by searching through all the original data—a tedious task when you have lots of data!

When looking at a data set, each observation may be considered as consisting of two parts—a stem and a leaf. To make a stem and leaf plot, each observed value must first be separated into its two parts:

- The stem is the first digit or digits;
- The leaf is the final digit of a value;
- Each stem can consist of any number of digits; but
- Each leaf can have only a single digit.

Example 2 – Making a stem and leaf plot

12, 23, 19, 6, 10, 7, 15, 25, 21, 12

Prepare a stem and leaf plot for these data.

Tip: The number 6 can be written as 06, which means that it has a stem of 0 and a leaf of 6.

The stem and leaf plot should look like this:

Table 2. Books read in a year by 10 students

Stem	Leaf
0	6 7
1	2 9 0 5 2
2	3 5 1

In Table 2:

- stem 0 represents the class interval 0 to 9;
- stem 1 represents the class interval 10 to 19; and
- stem 2 represents the class interval 20 to 29.

Usually, a stem and leaf plot is *ordered*, which simply means that the leaves are arranged in ascending order from left to right. Also, there is no need to separate the leaves (digits) with punctuation marks (commas or periods) since each leaf is always a single digit.

Using the data from Table 2, we made the **ordered** stem and leaf plot shown below:

Table 3. Books read in a year by 10 students

Stem	Leaf
0	6 7
1	0 2 2 5 9
2	1 3 5

Example 3 – Making an ordered stem and leaf plot

Fifteen people were asked how often they drove to work over 10 working days. The number of times each person drove was as follows:

5, 7, 9, 9, 3, 5, 1, 0, 0, 4, 3, 7, 2, 9, 8

Make an ordered stem and leaf plot for this table.

It should be drawn as follows:

Table 4. Number of drives to work in 10 days

Stem	Leaf
0	0 0 1 2 3 3 4 5 5 7 7 8 9 9 9

Splitting the stems

The organization of this stem and leaf plot does not give much information about the data. With only one stem, the leaves are overcrowded. If the leaves become too crowded, then it might be useful to split each stem into two or more components. Thus, an interval 0–9 can be split into two intervals of 0–4 and 5–9. Similarly, a 0–9 stem could be split into five intervals: 0–1, 2–3, 4–5, 6–7 and 8–9.

The stem and leaf plot should then look like this:

Table 5. Number of drives to work in 10 days

Stem	Leaf
$0^{(0)}$	0 0 1 2 3 3 4

$0^{(5)}$	5 5 7 7 8 9 9 9

Note: The stem $0^{(0)}$ means all the data within the interval 0–4. The stem $0^{(5)}$ means all the data within the interval 5–9.

Example 4 – Splitting the stems

Britney is a swimmer training for a competition. The number of 50-metre laps she swam each day for 30 days are as follows:

22, 21, 24, 19, 27, 28, 24, 25, 29, 28, 26, 31, 28, 27, 22, 39, 20, 10, 26, 24, 27, 28, 26, 28, 18, 32, 29, 25, 31, 27

1. Prepare an ordered stem and leaf plot. Make a brief comment on what it shows.
2. Redraw the stem and leaf plot by splitting the stems into five-unit intervals. Make a brief comment on what the new plot shows.

Answers

1. The observations range in value from 10 to 39, so the stem and leaf plot should have stems of 1, 2 and 3. The ordered stem and leaf plot is shown below:

Table 6. Laps swum by Britney in 30 days

Stem	Leaf
1	0 8 9
2	0 1 2 2 4 4 4 5 5 6 6 6 7 7 7 7 8 8 8 8 8 9 9
3	1 1 2 9

The stem and leaf plot shows that Britney usually swims between 20 and 29 laps in training each day.

2. Splitting the stems into five-unit intervals gives the following stem and leaf plot:

Table 7. Laps swum by Britney in 30 days

Stem	Leaf
$1^{(0)}$	0
$1^{(5)}$	8 9
$2^{(0)}$	0 1 2 2 4 4 4
$2^{(5)}$	5 5 6 6 6 7 7 7 7 8 8 8 8 8 9 9
$3^{(0)}$	1 1 2
$3^{(5)}$	9

Note: The stem $1^{(0)}$ means all data between 10 and 14, $1^{(5)}$ means all data between 15 and 19, and so on.

The revised stem and leaf plot shows that Britney usually swims between 25 and 29 laps in training each day. The values $1^{(0)}$ 0 = 10 and $3^{(5)}$ 9 = 39 could be considered *outliers*—a concept that will be described in the next section.

Example 5 – Splitting stems using decimal values

The weights (to the nearest tenth of a kilogram) of 30 students were measured and recorded as follows:

59.2, 61.5, 62.3, 61.4, 60.9, 59.8, 60.5, 59.0, 61.1, 60.7, 61.6, 56.3, 61.9, 65.7, 60.4, 58.9, 59.0, 61.2, 62.1, 61.4, 58.4, 60.8, 60.2, 62.7, 60.0, 59.3, 61.9, 61.7, 58.4, 62.2

Prepare an ordered stem and leaf plot for the data. Briefly comment on what the analysis shows.

Answer

In this case, the stems will be the whole number values and the leaves will be the decimal values. The data range from 56.3 to 65.7, so the stems should start at 56 and finish at 65.

Table 8. Weights of 30 students

Stem	Leaf
56	3
57	
58	4 4 9
59	0 0 2 3 8
60	0 2 4 5 7 8 9
61	1 2 4 4 5 6 7 9 9
62	1 2 3 7
63	
64	
65	7

In this example, it was not necessary to split stems because the leaves are not crowded on too few stems; nor was it necessary to round the values, since the range of values is not large. This stem and leaf plot reveals that the group with the highest number of observations recorded is the 61.0 to 61.9 group.

Outliers

An *outlier* is an extreme value of the data. It is an observation value that is significantly different from the rest of the data. There may be more than one outlier in a set of data.

Sometimes, outliers are significant pieces of information and should not be ignored. Other times, they occur because of an error or misinformation and should be ignored.

In the previous example, 56.3 and 65.7 could be considered outliers, since these two values are quite different from the other values.

By ignoring these two outliers, the previous example's stem and leaf plot could be redrawn as below:

Table 9. Weights of 30 students except for outliers

Stem	Leaf
58	4 4 9
59	0 0 2 3 8
60	0 2 4 5 7 8 9
61	1 2 4 4 5 6 7 9 9
62	1 2 3 7

When using a stem and leaf plot, spotting an outlier is often a matter of judgment. This is because, except when using box plots (explained in the section on box and whisker plots), there is no strict rule on how far removed a value must be from the rest of a data set to qualify as an outlier.

Features of distributions

When you assess the overall pattern of any distribution (which is the pattern formed by all values of a particular variable), look for these features:

- number of peaks
- general shape (skewed or symmetric)
- centre
- spread

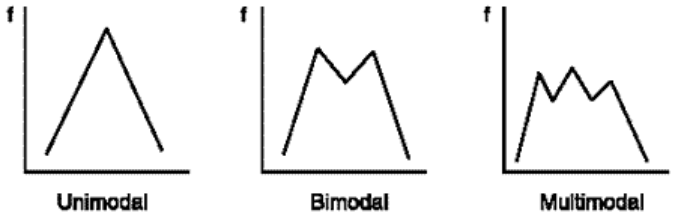
Number of peaks

Line graphs are useful because they readily reveal some characteristic of the data. (See the section on line graphs for details on this type of graph.)

The first characteristic that can be readily seen from a line graph is the number of high points or peaks the distribution has.

While most distributions that occur in statistical data have only one main peak *(unimodal),* other distributions may have two peaks *(bimodal)* or more than two peaks *(multimodal).*

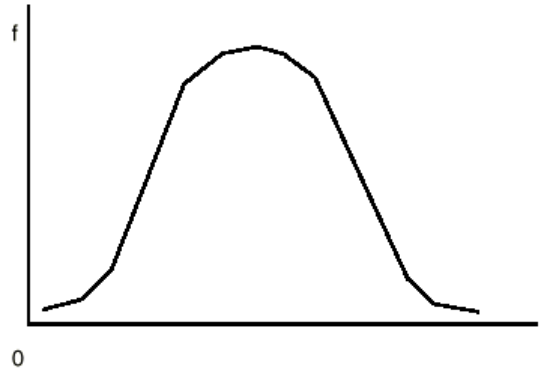
Examples of unimodal, bimodal and multimodal line graphs are shown below:



General shape

The second main feature of a distribution is the extent to which it is *symmetric.*

A *perfectly symmetric curve* is one in which both sides of the distribution would exactly match the other if the figure were folded over its central point. An example is shown below:

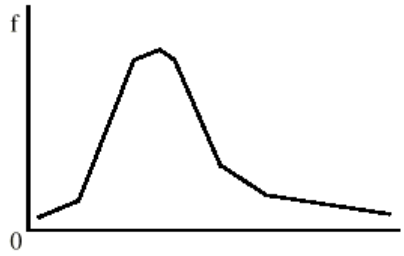


A symmetric, unimodal, bell-shaped distribution—a relatively common occurrence—is called *a normal distribution.*

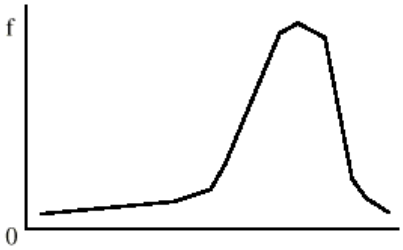
If the distribution is lop-sided, it is said to be *skewed.*

A distribution is said to be skewed to the right, or *positively skewed,* when most of the data are concentrated on the left of the distribution. Distributions with positive skews are more common than distributions with negative skews.

Income provides one example of a positively skewed distribution. Most people make under $40,000 a year, but some make quite a bit more, with a smaller number making many millions of dollars a year. Therefore, the positive (right) tail on the line graph for income extends out quite a long way, whereas the negative (left) skew tail stops at zero. The right tail clearly extends farther from the distribution's centre than the left tail, as shown below:



A distribution is said to be skewed to the left, or *negatively skewed,* if most of the data are concentrated on the right of the distribution. The left tail clearly extends farther from the distribution's centre than the right tail, as shown below:



Centre and spread

Locating the centre (*median*) of a distribution can be done by counting half the observations up from the smallest. Obviously, this method is impracticable for very large sets of data. A stem and leaf plot makes this easy, however, because the data are arranged in ascending order. The mean is another measure of central tendency. (See the chapter on central tendency for more detail.)

The amount of distribution spread and any large deviations from the general pattern (outliers) can be quickly spotted on a graph.

Using stem and leaf plots as graphs

A stem and leaf plot is a simple kind of graph that is made out of the numbers themselves. It is a means of displaying the main features of a distribution. If a stem and leaf plot is turned on its side, it will resemble a bar graph or histogram and provide similar visual information.

Example 6 – Using stem and leaf plots as graph

The results of 41 students' math tests (with a best possible score of 70) are recorded below:

31, 49, 19, 62, 50, 24, 45, 23, 51, 32, 48, 55, 60, 40, 35, 54, 26, 57, 37, 43, 65, 50, 55, 18, 53, 41, 50, 34, 67, 56, 44, 4, 54, 57, 39, 52, 45, 35, 51, 63, 42

1. Is the variable discrete or continuous? Explain.
2. Prepare an ordered stem and leaf plot for the data and briefly describe what it shows.
3. Are there any outliers? If so, which scores?
4. Look at the stem and leaf plot from the side. Describe the distribution's main features such as:
    a. number of peaks
    b. symmetry
    c. value at the centre of the distribution

Answers

1. A test score is a discrete variable. For example, it is not possible to have a test score of 35.74542341....
2. The lowest value is 4 and the highest is 67. Therefore, the stem and leaf plot that covers this range of values looks like this:

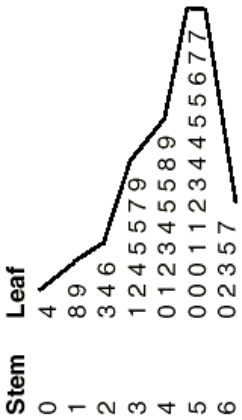
Table 10. Math scores of 41 students

Stem	Leaf
0	4
1	8 9
2	3 4 6
3	1 2 4 5 5 7 9
4	0 1 2 3 4 5 5 8 9
5	0 0 0 1 1 2 3 4 4 5 5 6 7 7
6	0 2 3 5 7

Note: The notation 2|4 represents *stem* 2 and *leaf* 4.

The stem and leaf plot reveals that most students scored in the interval between 50 and 59. The large number of students who obtained high results could mean that the test was too easy, that most students knew the material well, or a combination of both.

3. The result of 4 could be an outlier, since there is a large gap between this and the next result, 18.
4. If the stem and leaf plot is turned on its side, it will look like the following:



The distribution has a single peak within the 50–59 interval.

Although there are only 41 observations, the distribution shows that most data are clustered at the right. The left tail extends farther from the data centre than the right tail. Therefore, the distribution is skewed to the left or *negatively skewed.*

Since there are 41 observations, the distribution centre (the median value) will occur at the 21st observation. Counting 21 observations up from the smallest, the centre is 48. (Note that the same value would have been obtained if 21 observations were counted down from the highest observation.)

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 number of employees at Statistics Canada
    j. the number of brothers and sisters you have
    k. the distance between your house and school
    l. the number of pages in a dictionary

2. Without using any of the examples from question 1, give two examples of:
    a. a discrete variable
    b. a continuous variable

3. A telephone company surveyed 12 households to find out how many telephones there were per household.
    a. Copy the frequency distribution table below into your notebook and complete it using the following survey results:

    2, 5, 4, 3, 4, 3, 1, 3, 3, 2, 3, 4

Question 3a

Number of telephones (x)	Tally	Frequency (f)
1		
2		
3		
4		
5		

    b. Which result occurs most frequently?

4. 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. Are these discrete or continuous variables?
    b. Present these data in a frequency distribution table.
    c. Which result occurs most frequently?
    d. Set up a frequency distribution table including columns for the relative frequency and percentage frequency of the data.
    e. What conclusions can you draw from the tables? Explain.

5. A wind blew for 40 days. Its wind speeds, in knots, were recorded as follows:

    15, 22, 14, 12, 21, 34, 19, 11, 13, 0, 16, 4, 23, 8, 12, 18, 24, 17, 14, 3, 10, 12, 9, 15, 20, 5, 19, 13, 17, 11, 16, 19, 24, 12, 7, 14, 17, 10, 14, 23

    a. Are these discrete or continuous variables?
    b. Choose an appropriate class interval and present these data in a frequency distribution table.
    c. Which class interval occurs most frequently?
    d. Set up a frequency distribution table including columns for the relative frequency and percentage frequency of the data

e. What conclusions can be drawn from the tables? Explain.

6.

a. Prepare an ordered stem and leaf plot for the data in Exercise 5.

b. Do any outliers exist? If so, give a reason for their presence.

c. Describe the main features of the distribution:

i. number of peaks

ii. general shape

iii. approximate value at the centre of the distribution

7. Thirty people were surveyed to find out how often they went to the movie theatre in one year. The results are as follows:

21, 35, 27, 2, 18, 25, 10, 4, 43, 14, 29, 24, 15, 9, 26, 31, 41, 1, 28, 38, 40, 22, 37, 26, 19, 0, 33, 12, 16, 23

a. Copy the stem and leaf plot below into your notebook and complete it for the results.

Question 7a

Stem	Leaf
0	
1	
2	
3	
4	

b. Now, turn the plot into an ordered stem and leaf plot.

8. Assume the annual numbers of road fatalities from 1960 to 1992 were as follows:

10, 7, 8, 8, 17, 15, 17, 23, 14, 26, 31, 20, 32, 29, 31, 32, 38, 29, 30, 24, 30, 29, 26, 28, 37, 33, 32, 36, 32, 32, 26, 17, 20

a. Are these discrete or continuous variables?

b. Prepare an ordered stem and leaf plot of these data.

c. Expand the stem and leaf plot by using five-unit class intervals.

d. Do any outliers exist? If so, give a reason for their presence.

e. Describe the main features of the distribution:

i. number of peaks

ii. general shape

iii. approximate value at the centre of the distribution

9. From 1982 to 2002, the average minimum April temperature (Celsius) was recorded as follows:

6,1, 8,9, 6,9, 7,2, 7,0, 6,2, 5,7, 6,2, 6,8, 6,4, 6,8, 6,4, 7,6, 7,8, 7,3, 6,8, 8,8, 7,8, 8,1, 8,1, 7,9

a. Are these discrete or continuous variables?

b. Prepare an ordered stem and leaf plot for this data.

c. Is it necessary to expand the stem and leaf plot? Why or why not?

d. Do any outliers exist? If so, give a reason for their presence.

e. Describe the main features of the distribution.

i. number of peaks

ii. general shape

iii. approximate value at the centre of the distribution

10. Fifty staff members of a construction company were surveyed to find out what their weekly salary was to the nearest dollar. The results are as follows:

514, 476, 497, 511, 484, 513, 471, 470, 441, 466, 443, 481, 502, 528, 459, 548, 521, 517, 463, 478, 473, 514, 542, 519, 522, 523, 546, 487, 486, 473, 527, 470, 440, 564, 499, 523, 484, 463, 461, 437, 555, 525, 461, 539, 466, 470, 486, 490, 543, 519

a. Are these discrete or continuous variables?

b. Choose an appropriate class interval and present these data in a frequency distribution table.

c. Which class interval occurs most frequently?

d. Set up a frequency distribution table including columns for the relative frequency and percentage frequency of the data.

e. What conclusions can you draw from the tables? Explain.

f. Prepare an ordered stem and leaf plot for this data.

g. Do any outliers exist? If so, can you give a reason for their presence?

h. Looking at the stem and leaf plot, describe the main features of the distribution:

i. number of peaks

ii. general shape

iii. approximate value at the centre of the distribution

Class activities

1. Draw a straight line exactly 10 centimetres (cm) in length. Without measuring, place a mark where you estimate the halfway point to be. Now measure the line, and place a mark at the actual halfway point (5 cm). Measure the distance between your estimate and the actual halfway point. How many millimetres (mm) was your estimate short of the halfway point?

   a. Record this value in a table. Find out how far the rest of the class deviated from the halfway point and record these results.

   b. With these data, construct a frequency distribution table including columns for relative frequency and percentage frequency of the data.

   c. Which result occurs most frequently?

   d. Prepare a stem and leaf plot for this data.

   e. Do any outliers exist? If so, give a reason for their presence.

   f. Describe the main features of the distribution:

      i. number of peaks

      ii. general shape

      iii. approximate value at the centre of the distribution

   g. What conclusions can you draw from this analysis?

2. Ask your teacher for the class results (anonymous) from a recent test or assignment. Perform a detailed analysis on these data using the instructions described in a) to g) of Class activities 1. Briefly comment on:

   a. the class average of the test or assignment

   b. the ability of the class to understand the test or assignment questions

   c. the interest the class appears to have in the material tested

   Support each answer with evidence based on your analysis.

3. Throw one die 30 times. Using a frequency distribution table, record the result of each throw.

   a. Are these discrete or continuous variables?

   b. Set up a frequency distribution table including columns for the relative frequency and percentage frequency of the data.

   c. What result occurs most frequently?

   d. Did you expect any number to occur more often than the others? If so, why?

   e. Prepare a stem and leaf plot for these data.

   f. Do any outliers exist? If so, give a reason for their presence.

   g. Describe the main features of the distribution:

      i. number of peaks

      ii. general shape

      iii. approximate value at the centre of the distribution

   h. What conclusions can you draw from the analysis?

4. Create a table listing primary and secondary colours: red, blue, yellow, purple, green and orange. Include black, white, and grey as well. Do not include complementary shades of colours (blue-green, mauve, beige, etc.) in your table. Finally, label the last column "None of these colours".

   Then survey the teachers in your school to find out what colour car they drive. If a teacher responds with a colour that is not on your table, record their answer in the "None of these colours" column.

   a. Are these discrete or continuous variables?

   b. Set up a frequency distribution table including columns for the relative frequency and percentage frequency of the data.

   c. Determine which car colour is the most popular among the surveyed teachers. By what percentage is this colour more popular than the second most common colour?

   d. Why is it impossible to prepare a stem and leaf plot for this data?

   e. Why might a car manufacturer want this type of data analysis?

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. This is a continuous variable.

f. This is a discrete variable.

g. This is a continuous variable.

h. This is a continuous variable (this could also be considered discrete).

i. This is a discrete variable.

j. This is a discrete variable.

k. This is a continuous variable.

l. This is a discrete variable.

2. There are various answers to this question.

3.

a.

Answers to question 3a

Number of telephones (x)	Tally	Frequency (f)
1	I	1
2	II	2
3	𝍸	5
4	III	3
5	I	1
Total		**12**

b. The number 3 occurs the most frequently.

4.

a. These are discrete variables.

b.

Answers to question 4b

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

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

d.

Answers to question 4d

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**

5.

a. These are continuous variables.

b.

Answers to question 5b

Wind speed (x)	Tally	Frequency (f)
0 to < 5	III	3
5 to < 10	IIII	4
10 to < 15	𝍸 𝍸 IIII	14
15 to < 20	𝍸 𝍸 I	11
20 to < 25	𝍸 II	7
25 to < 30		0
30 to < 35	I	1
Total		**40**

c. The class interval that appears the most frequently is 10 –< 15.

d.

Answers to question 5d

Wind speed (x)	Frequency (f)	Relative frequency	Percentage frequency
0 to < 5	3	0.075	7.5
5 to < 10	4	0.100	10.0
10 to < 15	14	0.350	35.0
15 to < 20	11	0.275	27.5
20 to < 25	7	0.175	17.5
25 to < 30	0	0.000	0.0
30 to < 35	1	0.025	2.5
Total	**40**	**1.000**	**100.0**

e. The most commonly occurring wind speed is from 10 to less than 15 knots. Based on this 40-day sample, this wind speed has a 35% chance of occurring on any given day.

6.

a.

Answers to question 6a

Stem	Leaf
$0^{(0)}$	0 3 4
$0^{(5)}$	5 7 8 9
$1^{(0)}$	0 0 1 1 2 2 2 2 3 3 4 4 4 4
$1^{(5)}$	5 5 6 6 7 7 7 8 9 9 9
$2^{(0)}$	0 1 2 3 3 4 4

$2^{(5)}$	
$3^{(0)}$	4

b. The outlier in this exercise is the number 34. Possible explanations:

- a particularly windy or stormy day that took place during the 40 days
- there was a measurement error

c.

i. The distribution has only one main peak.

ii. The distribution is roughly asymmetrical. If the outlier were removed, it would be roughly skewed to the left. Generally, the distribution takes an irregular shape.

iii. The centre of distribution is 14 knots.

7.

a.

Answers to question 7a

Stem	Leaf
0	2 4 9 1 0
1	8 0 4 5 9 2 6
2	1 7 5 9 4 6 8 2 6 3
3	5 1 8 7 3
4	3 1 0

b.

Answers to question 7b

Stem	Leaf
0	0 1 2 4 9
1	0 2 4 5 6 8 9
2	1 2 3 4 5 6 6 7 8 9
3	1 3 5 7 8
4	0 1 3

8.

a. These are discrete variables.

b.

Answers to question 8b

Stem	Leaf
0	7 8 8
1	0 4 5 7 7 7
2	0 0 3 4 6 6 6 8 9 9 9
3	0 0 1 1 2 2 2 2 2 3 6 7 8

c.

Answers to question 8c

Stem	Leaf
$0^{(5)}$	7 8 8
$1^{(0)}$	0 4

$1^{(5)}$	5 7 7 7
$2^{(0)}$	0 0 3 4
$2^{(5)}$	6 6 6 8 9 9 9
$3^{(0)}$	0 0 1 1 2 2 2 2 3
$3^{(5)}$	6 7 8

d. No outliers exist in this exercise.

e.

i. The distribution has only one main peak.

ii. The general shape of the distribution is skewed to the left.

iii. The approximate value at the centre of the distribution is 28 road fatalities.

9.

a. These are continuous variables.

b.

Answers to question 9b

Stem	Leaf
5	7
6	1 2 2 4 4 8 8 8 9
7	0 2 3 6 8 8 9
7	1 1 8 9

c. No, it is not necessary to expand the stem and leaf plots because the stems are not overcrowded.

d. There are no outliers.

e.

i. The distribution has one main peak.

ii. The general shape of the distribution is roughly symmetric (although this is difficult to observe with such a small amount of data).

iii. The centre of distribution is 7 °C.

10.

a. These are discrete variables.

b.

Answers to question 10b

Weekly salary (x)	Tally	Frequency (f)
420 to < 440	I	1
440 to < 460	IIII	4
460 to < 480	𝍸 𝍸 IIII	14
480 to < 500	𝍸 IIII	9
500 to < 520	𝍸 III	8
520 to < 540	𝍸 III	8
540 to < 560	𝍸	5
560 to < 580	I	1
Total		**50**

c. The class interval that occurs the most frequently is $460–<480.

d.

Answers to question 10d

Weekly salary (x)	Frequency (f)	Relative frequency	Percentage frequency
420 to < 440	1	0.02	2
440 to < 460	4	0.08	8
460 to < 480	14	0.28	28
480 to < 500	9	0.18	18
500 to < 520	8	0.16	16
520 to < 540	8	0.16	16
540 to < 560	5	0.10	10
560 to < 580	1	0.02	2
Total	**50**	**1.00**	**100**

e. The most common salary group, representing 28% of the employees is between $460 and $480 a week based on this sample of 50 people. Only one staff member earned over $560 a week.

f.

Answers to question 10f

Stem	Leaf
43	7
44	0 1 3
45	9
46	1 1 3 3 6 6
47	0 0 0 1 3 3 6 8
48	1 4 4 6 6 7
49	0 7 9
50	2
51	1 3 4 4 7 9 9
52	2 3 3 5 7 8
53	9
54	2 3 6 8
55	5
56	4

g. There are no outliers. If there were outliers a possible explanation for them could be:

- these people may have been managers or directors and thus on higher salaries
- these people may have given misleading responses.

h.

i. The distribution has many peaks which indicated the possibility of being bimodal.

ii. The distribution is neither symmetrical nor is it skewed.

iii. The centre of the distribution is between $487 and $490.

Graph types

Graphs 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.

Using graphs

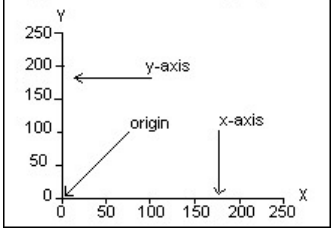
What is a graph?

A graph is a visual representation of a relationship between, but not restricted to, two variables. A graph generally takes the form of a one- or two-dimensional figure such as a scatterplot. Although, there are three-dimensional graphs available, they are usually considered too complex to understand easily.

A graph commonly consists of two axes called the x-axis (horizontal) and y-axis (vertical). Each axis corresponds to one variable. The axes are labelled with different names, such as *Price* and *Quantity*.

The place where the two axes intersect is called the origin. The origin is also identified as the point (0,0).

Figure 1. Parts of a graph



A point on a graph represents a relationship. Each point is defined by a pair of numbers containing two co-ordinates (x and y). A co-ordinate is one of a set of numbers used to identify the location of a point on a graph.

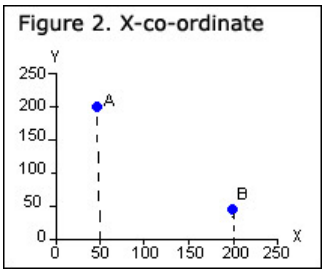
In the following section, you will learn how to determine both co-ordinates for any given point, and to correctly label the co-ordinates of a point.

Identifying the x-co-ordinate

The x-co-ordinate of a point is the value that tells you how far the point is from the origin on the (horizontal) x-axis. In order to find the x-co-ordinate of a point on any graph, draw a straight line from the point to intersect at a right angle with the x-axis. The number where the line intersects with the x-axis is the value of the x-co-ordinate.

Figure 2 is a graph with two points, A and B. Identify the x-co-ordinate of points A and B.

Figure 2. X-co-ordinate



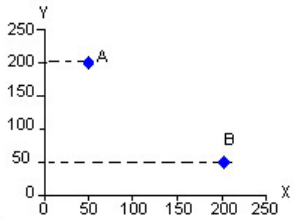
Answer: The x-co-ordinate of point A is 50, and the x-co-ordinate of point B is 200.

Identifying the y-co-ordinate

The y-co-ordinate of a point is the value that tells you how far away the point is from the origin on the vertical or y-axis. To find the y-co-ordinate of a point on a graph, draw a straight line from the point to intersect at a right angle with the y-axis. The number where the line intersects the y-axis is the value of the y-co-ordinate.

Identify the y-co-ordinate for point A and point B on Figure 3.

Figure 3. Y-co-ordinate

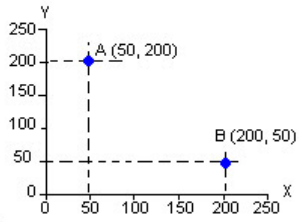


Answer: The y-co-ordinate of point A is 200, and the y-co-ordinate of point B is 50.

Identifying points on a graph

Once you have determined the co-ordinates of a point, you can label the points using ordered pair notation. This notation is simple—points are identified by stating their co-ordinates in the form of (x, y). Note that you must plot the x-co-ordinate first as in Figure 2. The x- and y-co-ordinates for each of points A and B are identified in Figure 4 below.

Figure 4. Plotting co-ordinates using ordered pair notation

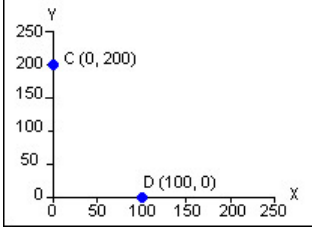


- The x-co-ordinate of point A is 50 and the y-co-ordinate of point A is 200. The co-ordinates of point A are therefore (50, 200).
- The x-co-ordinate of point B is 200 and the y-co-ordinate of point B is 50. The co-ordinates of point B are therefore (200, 50).

Points on the axes

If a point falls on an axis, you do not need to draw lines to determine the co-ordinates of the point. In Figure 5 below, point C lies on the y-axis and point D lies on the x-axis. When a point lies on an axis, one of its co-ordinates must be 0.

Figure 5. Points on the axes



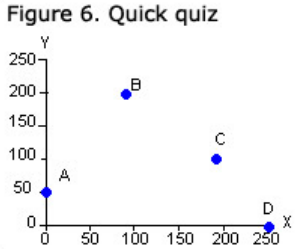
- Point C lies on the y-axis and has an x-co-ordinate of 0. When you move along the y-axis to find the y-co-ordinate, the point is 200 from the origin. The co-ordinates of point C are therefore (0, 200).
- Point D lies on the x-axis and has a y-co-ordinate of 0. If you move along the x-axis to find the co-ordinate, the point is 100 from the origin. The co-ordinates of point D are therefore (100, 0).

Quick quiz!

Answer the following questions using Figure 6 below.

- Which points intersect with the y-axis?
- Which point would be labelled with the ordered pair notation of (100, 200)?
- Which points have a y-co-ordinate of 100?

Figure 6. Quick quiz



Answers; 1. Point A 2. Point B 3. Point C

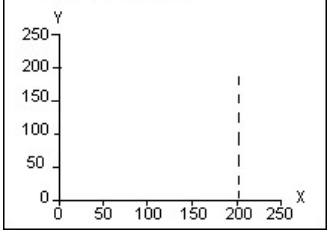
Plotting points on a graph

There are times when you will be given the coordinates of a point and will need to find its location on a graph. This process is often referred to as plotting a point. The process for plotting a point is shown below.

Plot the point (200, 150) using the following step-by-step approach.

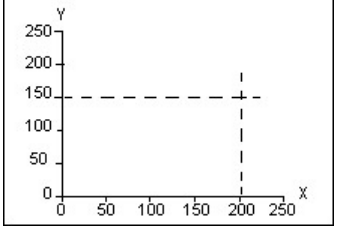
Step 1
First, draw a perpendicular line extending out from the x-axis at the x-co-ordinate of the point. In the example, the x-co-ordinate is at 200.

Figure 7. Step 1



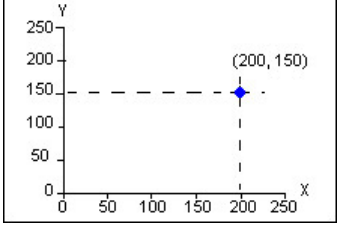
Step 2
Then, draw a perpendicular line extending out from the y-axis at the y-co-ordinate of the point, the y-co-ordinate is at 150.

Figure 8. Step 2



Step 3
Finally, draw a dot where the two lines intersect. This is the point we are plotting (200, 150).

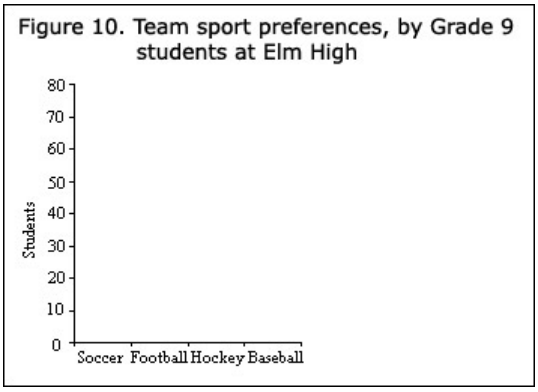
Figure 9. Step 3



Deciding on a scale

The scale of a graph is very important. It is determined by the data for each axis, and should be measured accordingly.

Figure 10. Team sport preferences, by Grade 9 students at Elm High



A survey was conducted of the Grade 9 students at Elm High. The students were asked which of the following four team sports they preferred.

The results were:

1. Soccer – 45 students
2. Football – 55 students
3. Hockey – 75 students
4. Baseball – 25 students

In Figure 10, these four preference categories have been placed on the x-axis, each representing the grouped data collected. Because the categories are nominal (names, not numbers) and describe qualitative (not quantitative) distinctions, the groups can be placed in any order on the axis.

On the y-axis, the data values range from 0 to 80 students. As mentioned earlier, your origin should be located at 0 where the x-axis and y-axis meet. Since the largest group of students by sport preference is 75, then it would be appropriate to end the scale at 80, resulting in a scale that ranges from 0 to 80. Depending on how the scale is arranged, the graph will not change, but its visual appearance might be altered.

The interval of the scale is the amount of space along the axis from one mark to the next. If the range of the scale is small, the general rule is to take the range of the scale and divide it by 10. Make this your interval. For ranges that are larger, the interval is typically 5, 10, 100, 500, 1,000, etc. Use numbers that divide evenly into 100, 1,000 or their multiples in order to provide a graph that is easy to understand.

In this case, if you take 80 and divide it by 5, you will get 16. However, it might be better to use 10 because it is easier to analyse. This provides a scale that is smaller, but still easy to use.

Rules for good 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 graphs
- pictographs
- circle graphs/pie charts
- line graphs
- scatterplots
- histograms

Knowing what type of graph to use with what type of information is crucial. Depending on the nature of the data some graphs are more appropriate than others. For example, categorical data like favorite school subjects are best displayed in a bar graph or circle graph while continuous numeric data such as height are illustrated by a line graph or histogram. For more information on appropriate graph types, see "Types of data" in *Teacher's Guide to Data Discovery*.

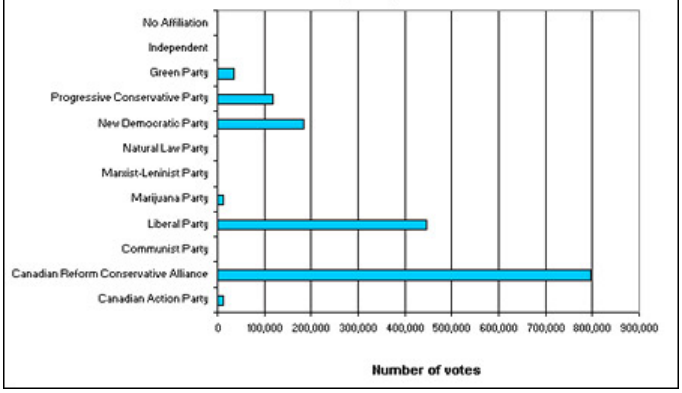
When is it not appropriate to use a graph?

A graph is not always the most appropriate tool to present information. Sometimes text or a data table can provide a better explanation to the readers—and save you considerable time and effort.

You might want to reconsider the use of a graph when

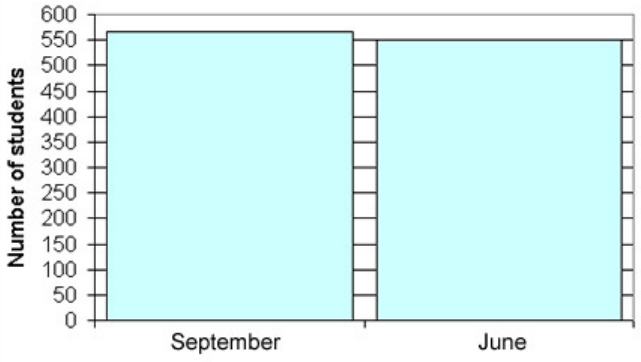
- the data are very dispersed

Figure 11. Division of votes for the major political parties, in a federal election, Anytowne



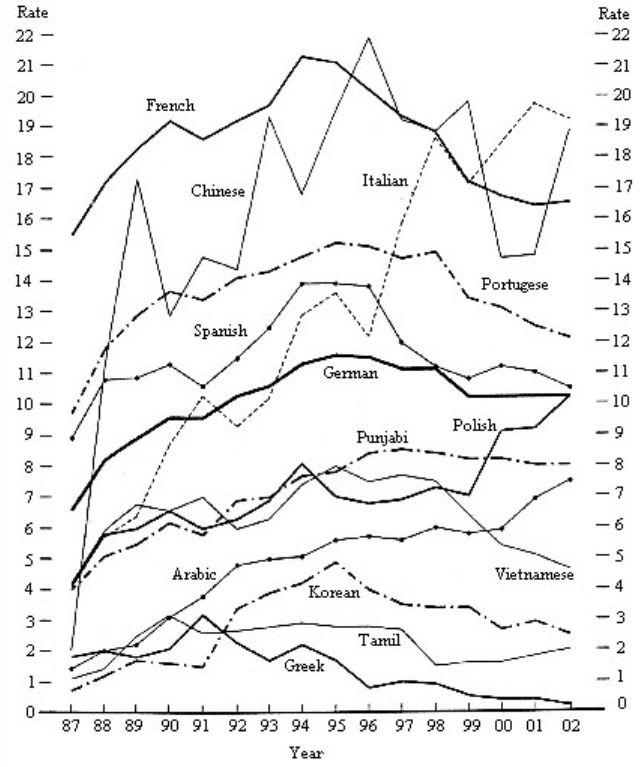
- there are too few data (only one, two or three data points)

Figure 12. Number of students enrolled in Greenfield Secondary School



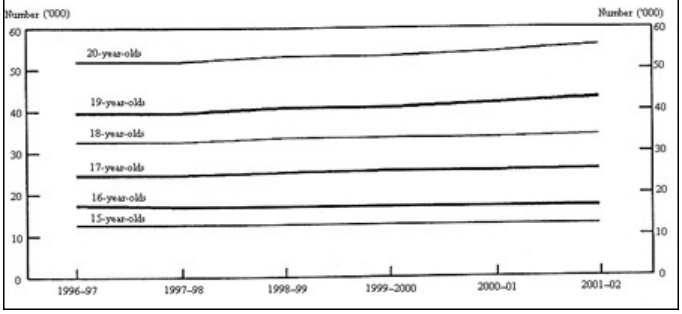
- the data are very numerous

Figure 13. Number of students taking English as a second language at West High School, by first language spoken, 1987 to 2002



- the data show little or no variations

Figure 14. Number of young adults who exercise at least once weekly, by age, 1996 to 2002



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:

   a. Who is your target audience?
   b. What do they know about the issue?
   c. What do they expect to see?
   d. What do they want to know?
   e. What will they do with 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:

   a. What do the data show?
   b. Is there more than one main message?
   c. What aspect of the message(s) should be highlighted?
   d. Can all the messages be displayed in the same graphic?

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:

Use appropriate terms to describe your graph

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 in data	increase with, decrease with, vary with, despite, correspond to, relate to

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

   a. circle graph/pie chart (description of components)
   b. horizontal bar graph (comparison of items and relationships, time series)
   c. vertical bar graph (comparison of items and relationships, time series, frequency distribution)
   d. line graph (time series and frequency distribution)
   e. scatterplot (analysis of relationships)

Bar graphs

A bar graph may be either horizontal or vertical. The important point to note about bar graphs is their bar length or height—the greater their length or height, the greater their value.

Bar graphs 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 graphs usually present categorical and numeric variables grouped in class intervals. They consist of an axis and a series or labeled horizontal or vertical bars. The bars depict frequencies of different values of a variable or simply the different values themselves. The numbers on the x-axis of a bar graph or the y-axis of a column graph are called the scale.

When developing bar graphs, 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.

There are three types of graphs used to display time series data:

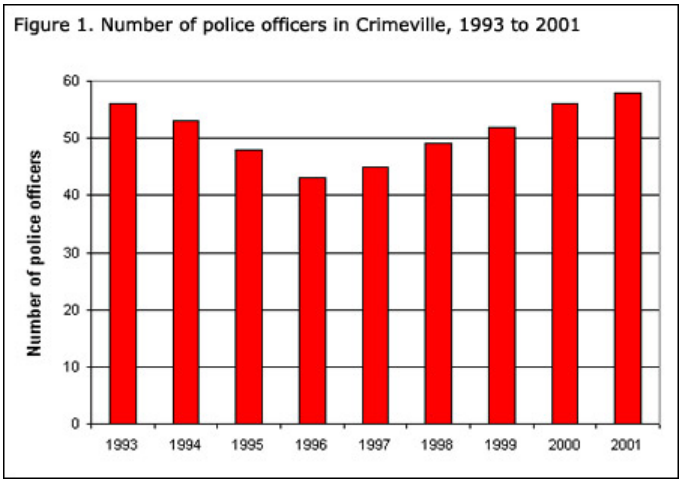
- horizontal bar graphs,
- vertical bar graphs and
- line graphs.

All three of these types of graphs work well when you need to compare values. However, in general, data comparisons are best represented vertically.

Example 1 – Vertical bar graphs

Bar graphs should be used when you are showing segments of information. From the information given in the section on graph types, you will know that vertical bar graphs are particularly useful for time series data. The space for labels on the x-axis is small, but ideal for years, minutes, hours or months. At a glance you can see from the graph that the scales for both the x- and y-axis increase as they get farther away from the origin. Figure 1 below shows the number of police officers in Crimeville from 1993 to 2001.

Figure 1. Number of police officers in Crimeville, 1993 to 2001

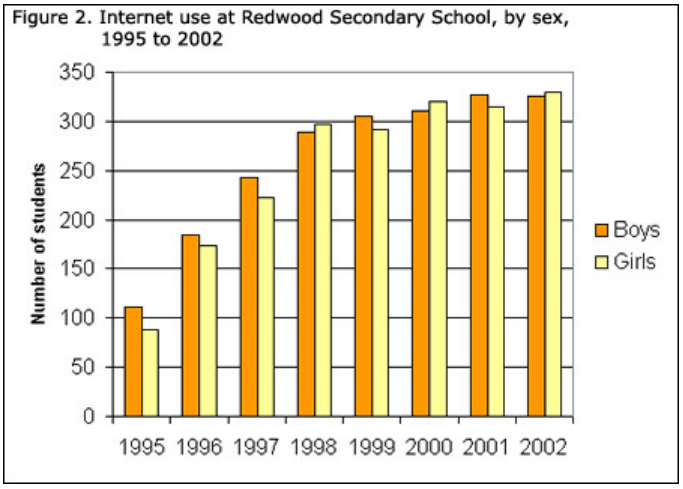


In Figure 1 you can see that the number of police officers decreased from 1993 to 1996, but started increasing again in 1996. The graph also makes it easy to compare or contrast the number of police officers for any combination of years. For example, in 2001 there were nine more police officers than in 1998.

The double (or group) vertical bar graph is another effective means of comparing sets of data about the same places or items. This type of vertical bar graph gives two or more pieces of information for each item on the x-axis instead of just one as in Figure 1. This allows you to make direct comparisons on the same graph by age group, sex, race, or anything else you wish to compare. However, if a group vertical bar graph has too many series of data, the graph becomes cluttered and it can be confusing to read.

Figure 2, a double vertical bar graph, compares two series of data: the numbers of boys and girls using the Internet at Redwood Secondary School from 1995 to 2002. One bar represents the number of boys who use the Internet and the other bar represents the girls.

Figure 2. Internet use at Redwood Secondary School, by sex, 1995 to 2002



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

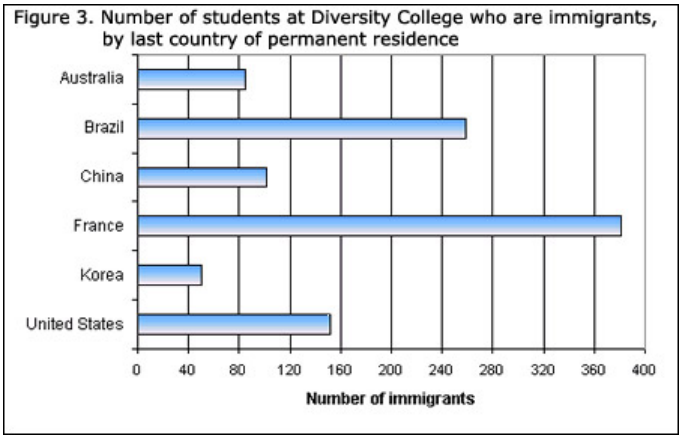
Example 2 – Horizontal bar graphs

The horizontal bar graph uses the y-axis (vertical line) for labelling. There is more room to fit text labels for categorical variables on the y-axis.

Figure 3 shows the number of students at Diversity College who are immigrants by their last country of permanent residence. The graph shows that 100 students immigrated from China, 380 from France, and 260 from Brazil.

A horizontal bar graph has been used to show a comparison of these data. This graph is the best method to present this type of information because the labels (in this case, the countries' names) are too long to appear clearly on the x-axis.

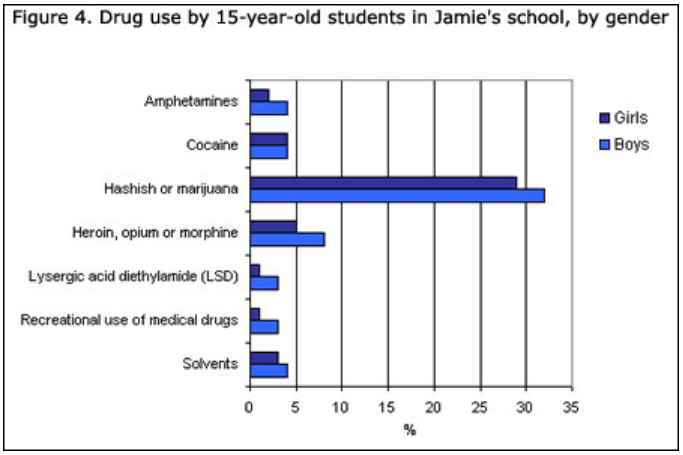
Figure 3. Number of students at Diversity College who are immigrants, by last country of permanent residence



A double or group horizontal bar graph is similar to a double or group vertical bar graph, and it would be used when the labels are too long to fit on the x-axis.

In Figure 4, more than one piece of information is being delivered to the audience: drug use by 15-year-old boys is being compared with drug use by 15-year-old girls at Jamie's school. Having both pieces of information on the same graph makes it easier to compare. The graph indicates that 32% of boys and 29% of girls have used hashish or marijuana, and 3% of boys and 1% of girls have tried LSD. The graph also shows that the same percent of boys and girls (4%) have used cocaine.

Figure 4. Drug use by 15-year-old students in Jamie's school, by gender



Example 3 – Comparing several places or items

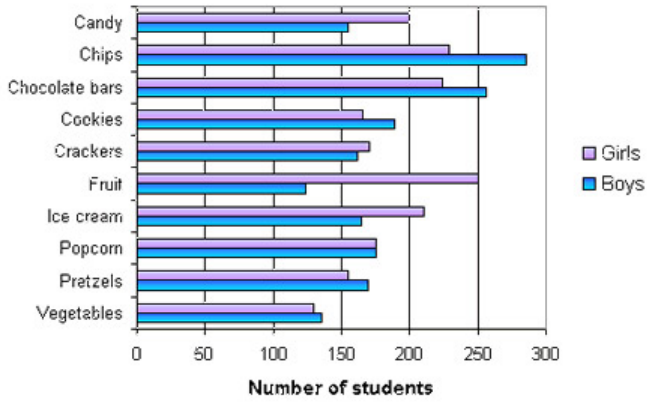
Figure 5 is an example of a double horizontal bar graph. Hillary sampled an equal number of boys and girls at her high school and asked them to pick the one snack food they liked the most from the following list:

- popcorn
- chips
- chocolate bars
- crackers
- pretzels
- cookies
- ice cream
- fruit
- candy
- vegetables.

She created a graph to display the results of her survey. Examine Figure 5, and answer the following questions:

1. What comparison does this graph show?
2. Which snack food was least preferred by girls?
3. Which snack food was preferred by substantially more boys than girls?
4. Which snack foods were preferred by more girls than boys?
5. Which snack food was preferred equally by both boys and girls?

Figure 5. Preferred snack choices of students at Hillary's high school



Answers

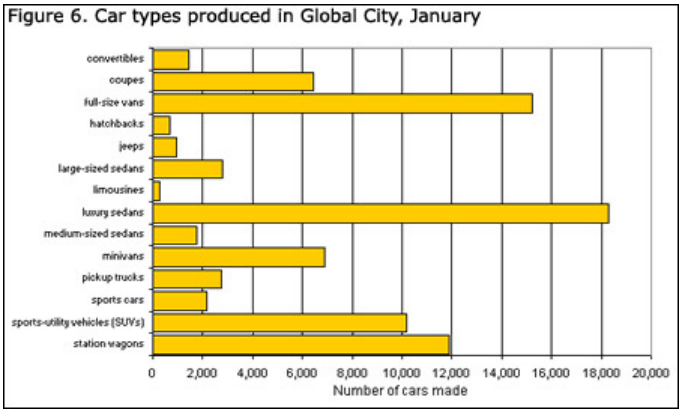
1. The graph shows a comparison of snack food preferences by sex.
2. Vegetables were the snack food least preferred by girls.
3. A substantial number of more boys than girls preferred chips.
4. Girls preferred candy, crackers, fruit and ice cream more than boys did.
5. The same number of boys and girls preferred popcorn as their snack food choice.

Example 4 – Inappropriate use of bar graphs

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

A horizontal bar graph may be more effective than a line graph when there are fewer time periods or segments of data. If you want to compare more than 9 or 10 items, use a line graph instead. Figure 6 is an example of when a line graph should be used instead of a horizontal bar graph.

Figure 6. Car types produced in Global City, January



Example 5 – Other bar graphs

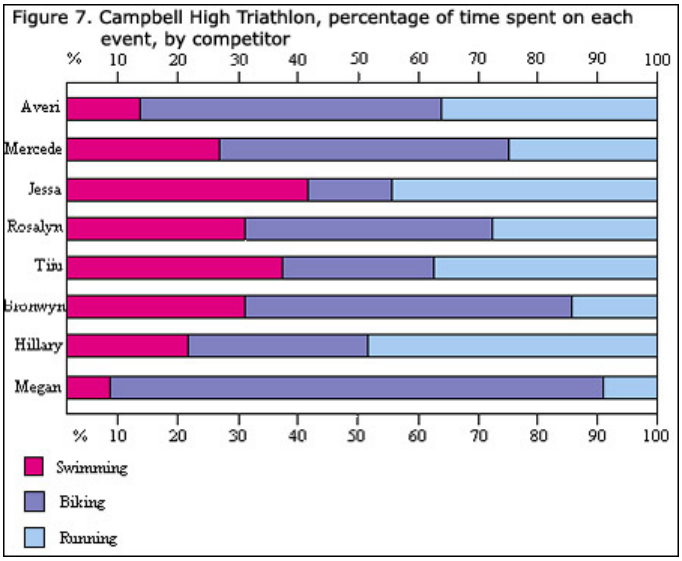
There are several other types of bar graphs that you may encounter. The population pyramid is a special application of a double bar graph. The following examples are rarely used, but can be useful if used correctly.

Stacked bar graphs

The stacked bar graph is a preliminary data analysis tool used to show segments of totals. Statistics Canada rarely uses them, despite the fact that stacked bar graphs can convey a lot of information. The stacked bar graph can be very difficult to analyse if too many items are in each stack. It can contrast values, but not necessarily in the simplest manner.

In Figure 7, it is not difficult to analyse 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 graph representing a decathlon (with 10 events) the data would have been significantly harder to analyse.

Figure 7. Campbell High Triathlon, percentage of time spent on each event, by competitor



Another reason that these graphs are rarely used is that they can represent a picture other than the one that was intended. In the example above, it may have taken Bronwyn two hours to finish the triathlon, and Rosalyn three hours, but they spent almost the same percentage of time on each event. Both women spent 30% of their times swimming, but whereas Rosalyn spent 54 minutes swimming, Bronwyn spent 36 minutes swimming. In other words, this graph does not tell you anything about their ranking, only what percentage of their individual race times were spent on each event. This can be misleading for someone who does not read the graph carefully.

Horizontal, vertical and stacked bar graph guidelines

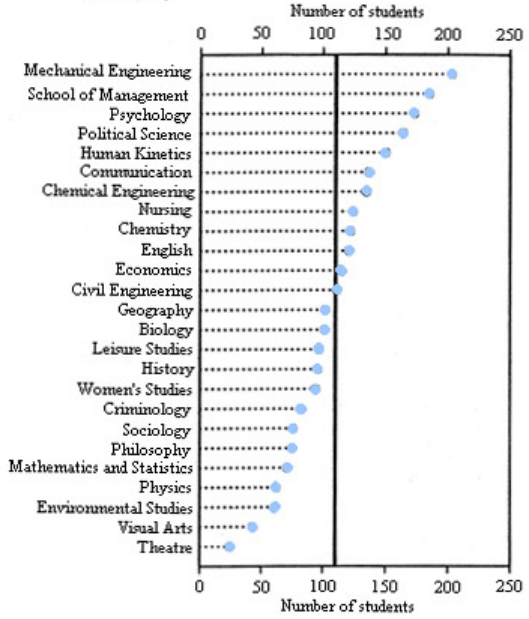
You should keep the following guidelines in mind when creating your own bar graphs:

- Make bars and columns wider than the space between them.
- Do not allow grid lines to pass through columns or bars.
- Use a single font type on a graph. Try to maintain a consistent font style from graph to graph in a single presentation or document. Simple sans-serif fonts are preferable.
- Order your shade pattern from darkest to lightest on stacked bar graphs.
- Avoid garish colours or patterns.

Dot graphs

A dot graph is one of the simplest ways to represent information pictorially, yet it is the graph that is least used. Figure 8 is an example of a dot graph. As you can see, the message and the information behind the graph are delivered quickly and easily to the reader.

Figure 8. Rates of enrolment by program, Academia University



Pictographs

A pictograph uses picture symbols to convey the meaning of statistical information. Pictographs should be used carefully because the graphs may, either accidentally or deliberately, misrepresent the data. This is why a graph should be visually accurate.

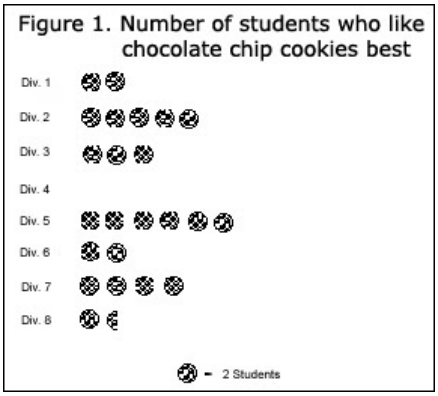


Figure 1 shows a scale that represents the number of elementary students who prefer chocolate chip cookies. This type of pictograph shows how a symbol can be designed to represent data. One cookie symbol represents two students, and a half-cookie image is used to represent one student. These data could have been easily presented in a histogram where the figure is expressed using a scale rather than a symbol.

Now let us look at another example of a pictograph.

Figure 2 shows how the Canadian dollar shrank to a value of 46.17 cents over 20 years because of inflation. This information means the value of the 2000 Canadian dollar was worth less than half as much as it was in 1980! What is the problem with the depiction of statistics in this pictograph?

The size or area (total surface) of the dollars coin (loonie) graphic is misleading. The dollar value differences represented are exaggerated by the pictures. The graphics should reflect the actual purchasing power of the dollar of the year in question. Since 46 cents is just under half of one dollar, the 2000 loonie should appear to be just under half the size of the 1980 loonie. Instead of being one-quarter of the size of the 1980 loonie, the 2000 loonie should be about twice as big as is shown.

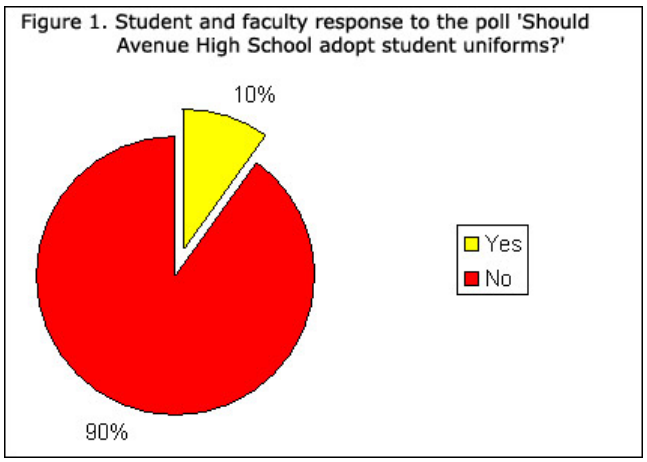
You may argue that people do not notice this misrepresentation when they look at a pictograph such as this one, and thus it is not particularly important. The fact is that subconsciously many people interpret the Canadian dollar to have lost far more of its value than it has in reality. Since many people 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.

Circle graphs/pie charts

A circle graph/pie chart is a way of summarizing a set of categorical 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.

Circle graphs/pie charts usually show the component parts of a whole. Often you will see a segment of the drawing separated from the rest of the pie in order to emphasize an important piece of information.



The circle graph/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 adopt school uniforms. This point is clearly emphasized by its visual separation from the rest of the pie.

The use of the circle graph/pie chart is quite popular, as the circle provides a visual concept of the whole (100%). Circle graphs/pie charts are also one of the most commonly used charts because they are simple to use. Despite its popularity, circle graphs/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, circle graphs/pie charts are not useful when the values of each component are similar because it is difficult to see the differences between slice sizes.

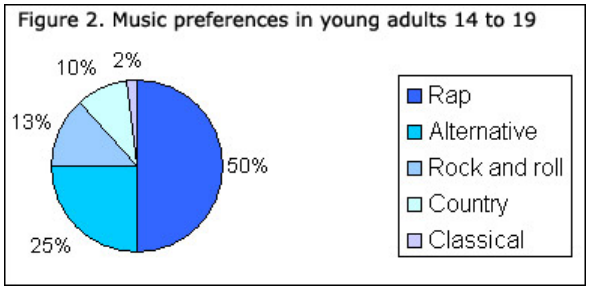
A circle graph/pie chart uses percentages to compare information. Percentages are used because they are the easiest way to represent a whole. The whole is equal to 100%. For example, if you spend 7 hours at school and 55 minutes of that time is spent eating lunch, then 13.1% of your school day was spent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

percent ÷ 100 x 360 degrees = the number of degrees

This ratio works because the total percent of the circle graph/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 circle graph/pie chart

A circle graph/pie chart is constructed by converting the share of each component into a percentage of 360 degrees. In Figure 2, music preferences in 14- to 19-year-olds are clearly shown.



The circle graph/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 circle graph/pie chart, ensure that the segments are ordered by size (largest to smallest) and in a clockwise direction.*

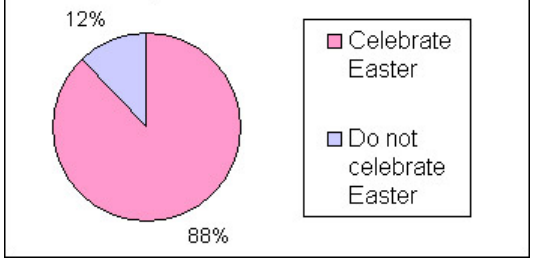
In order to reproduce this circle graph/pie chart, follow this step-by-step approach:

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

1. Draw a circle with your protractor.
2. 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.
3. 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.

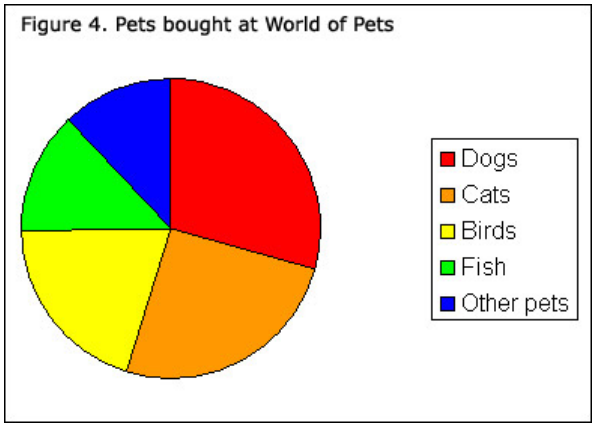
Labeling the segments with percentage values often makes it easier to tell quickly which segment is bigger. Whenever possible, the percentage and the category label should be indicated beside their corresponding segments. This way, users do not have to constantly look back at the legend in order to identify what category each colour represents.

Figure 3. Percentage of students in Mr. Paul's World Religions class who celebrate Easter



The circle graph/pie chart above conveys a clear message to the user—that 88% 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. Unfortunately, the category labels are too long to fit beside the pie segments, so they had to be placed in the legend. Ideally, these labels would also accompany the pie segments.

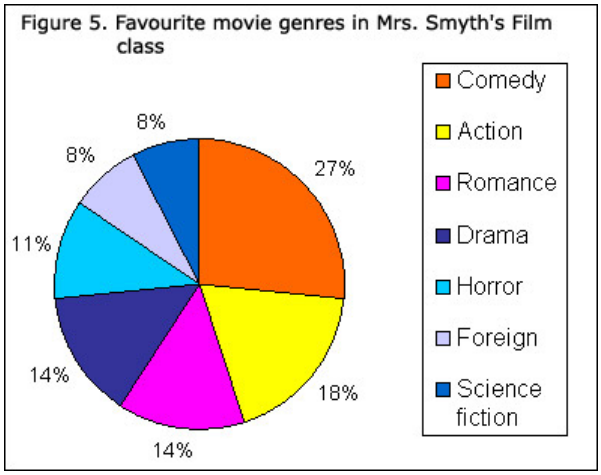
Figure 4. Pets bought at World of Pets



It is more difficult to understand the message behind Figure 4 because there are no percentage figures given for each slice of the pie. This is why it is important to label the slices with actual values.

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.

Figure 5. Favourite movie genres in Mrs. Smyth's Film class



In the circle graph/pie chart above, the legend is formatted properly and the percentages are included for each of the pie segments. However, there are too many items in the circle graph/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 a another graph to display the information. Figure 5 would certainly be easier to read as a bar graph.

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

Circle graphs/pie charts versus bar graphs

When displaying statistical information, refrain from using more than one circle graph/pie chart for each figure.

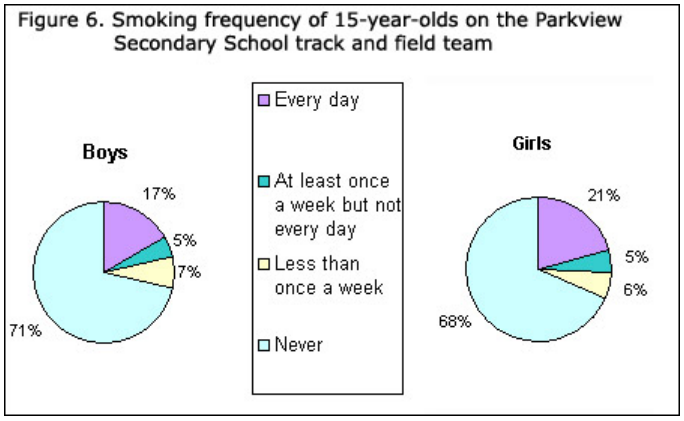
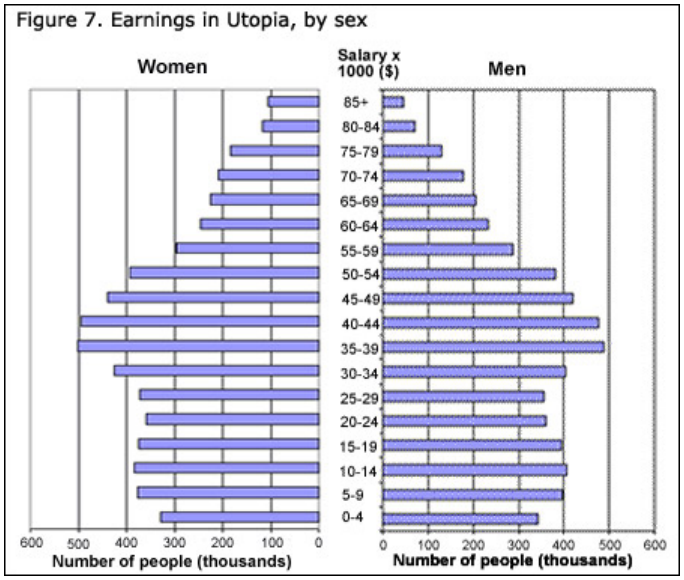


Figure 6 shows two circle graphs/pie charts side-by-side, where a split bar graph (two bar graphs back-to-back) would have shown the information more clearly. A user might find it difficult to compare a segment from one circle graph/pie chart to the corresponding segment of the other circle graph/pie chart. However, in a split bar graph, these segments become bars which are lined up back-to-back, making it much easier to make comparisons.

Figure 7 shows how a split bar graph would be a better choice for displaying information than a double circle graph/pie chart. The key point in preparing this type of graph is to ensure that you are using the same scale for both sides of the bar graph. You'll notice that the information is much clearer in Figure 7 than in Figure 6.



Line graphs

Line graphs, 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 graphs are easy to create.

A line graph 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. For information on the shapes of line graphs, see the Organizing data chapter.

Line graphs 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 graph usually indicates quantity (e.g., dollars, litres) or percentage, while the horizontal x-axis often measures units of time. As a result, the line graph 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 graphs are able to show relationships more clearly than tables do. Line graphs can also depict multiple series and hence are usually the best candidate for time series data and frequency distribution.

Bar and column graphs and line graphs share a similar purpose. The column graph, however, reveals a change in magnitude, whereas the line graph is used to show a change in direction.

In summary, line graphs

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

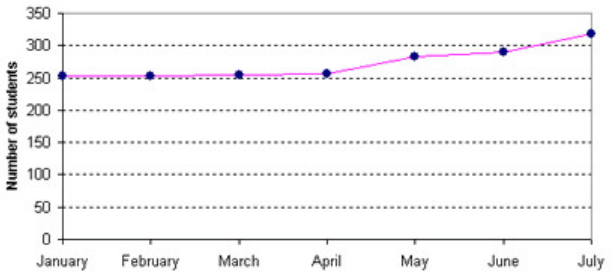
Graphs can give a distorted image of the data. If inconsistent scales on the axes of a line graph force data to appear in a certain way, then a graph can even reveal a trend that is entirely different from the one intended. This means that the intervals between adjacent points along the axis may be dissimilar,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

Example 1 – Plotting a trend over time

Figure 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.

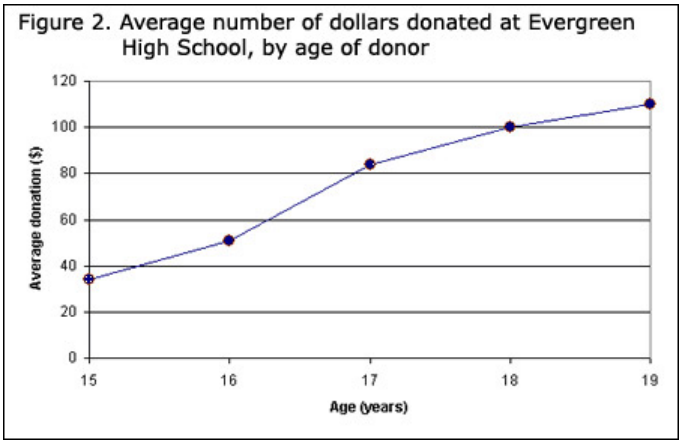
The number of students participating in the labour force was 252 in January, 252 in February, 255 in March, 256 in April, 282 in May, 290 in June and 319 in July. When examined further, the graph indicates that the labour force participation of these students was at a plateau for the first four months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

Figure 1. Labour force participation in Andrew's high school



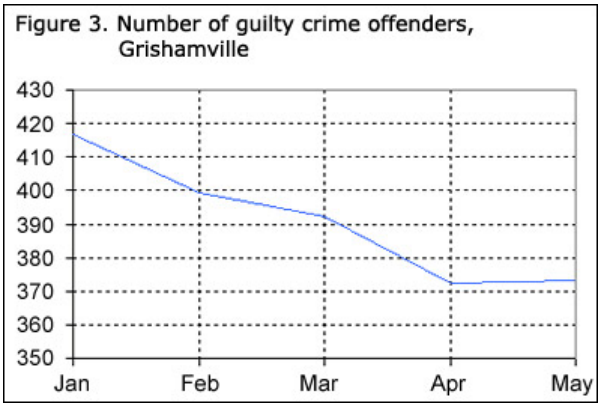
Example 2 – Comparing two related variables

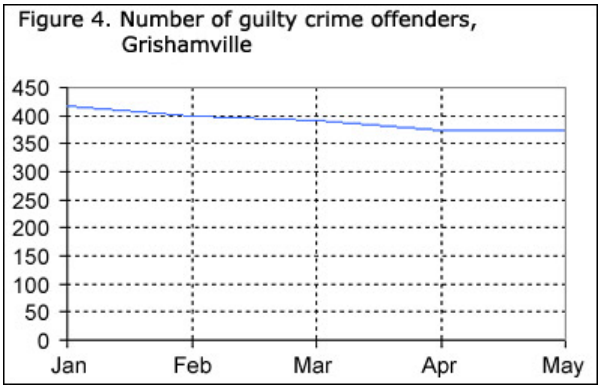
Figure 2 is a single line graph comparing two items; in this instance, time is not a factor. The graph compares the number of dollars donated by the age of the donors. According to the trend in the graph, the older the donor, the more money he or she donates. The 17-year-old donors donate, on average, $84. For the 19-year-olds, the average donation increased by $26 to make the average donation of that age group $110.



Example 3 – Using correct scale

When drawing a line, it is important that you use the correct scale. Otherwise, the line's shape can give readers the wrong impression about the data. Compare Figure 3 with Figure 4:





Using a scale of 350 to 430 (Figure 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 and does not relate it to the bigger picture. However, choosing a scale of 0 to 450 (Figure 4) better displays how small the decline in the number of guilty crime offenders really was.

Example 4 – Multiple line graphs

A multiple line graph can effectively compare similar items over the same period of time (Figure 5).

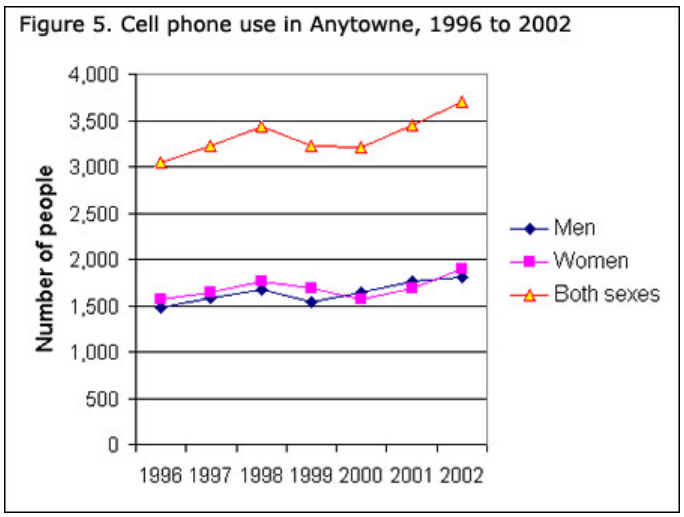


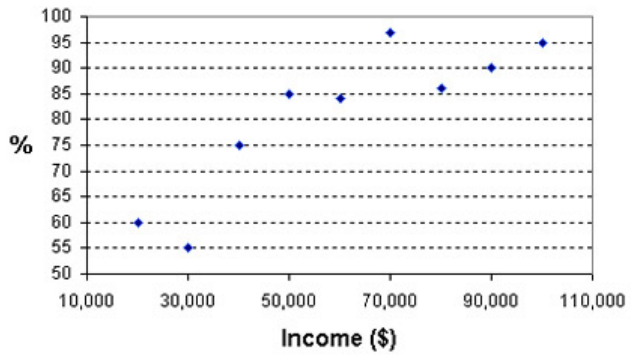
Figure 5 is an example of a very good graph. The message is clearly stated in the title, and each of the line graphs is properly labelled. It is easy to see from this graph that the total cell phone use has been rising steadily since 1996, except for a two-year period (1999 and 2000) where the numbers drop slightly. The pattern of use for women and men seems to be quite similar with very small discrepancies between them.

Scatterplots

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

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 (see Figure 1 below).

Figure 1. Car ownership in Anytowne, by household income



Car ownership increases as the household income increases, showing that there is a positive relationship between these two variables.

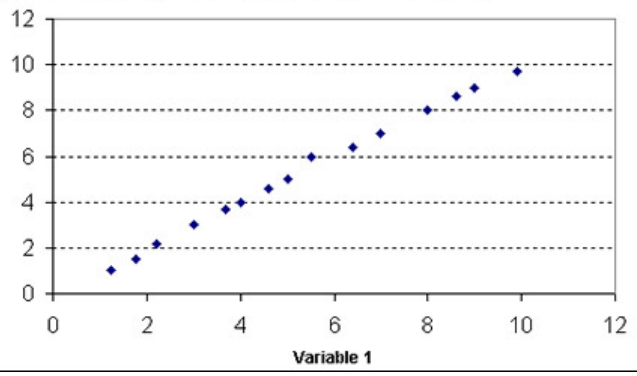
The pattern of the data points on the scatterplot reveals the relationship between the variables. Scatterplots can illustrate various patterns and relationships, such as:

- data correlation
- positive or direct relationships between variables
- negative or inverse relationships between variables
- scattered data points
- non-linear patterns
- spread of data
- outliers.

Data correlation

When the data points form a straight line on the graph, the linear relationship between the variables is stronger and the correlation is higher (Figure 2).

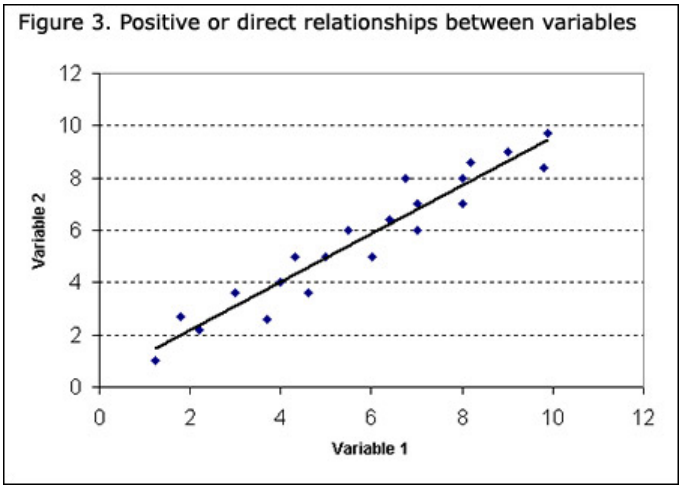
Figure 2. Strong linear relationship of variables



Positive or direct relationships

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 positive or direct (Figure 3). An increase in the value of x is more likely associated with an increase in the value of y. The closer the points are to the line, the stronger the relationship.

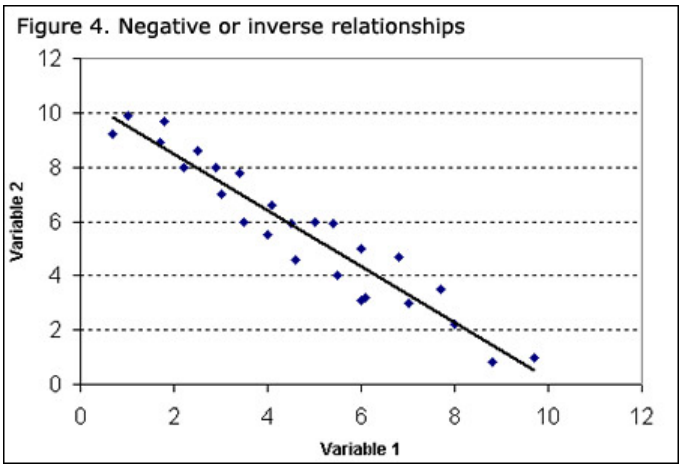
Figure 3. Positive or direct relationships between variables



Negative or inverse relationships

If the points tend to cluster around a line that runs from the upper left to lower right of the graph, then the relationship between the two variables is negative or inverse (Figure 4).

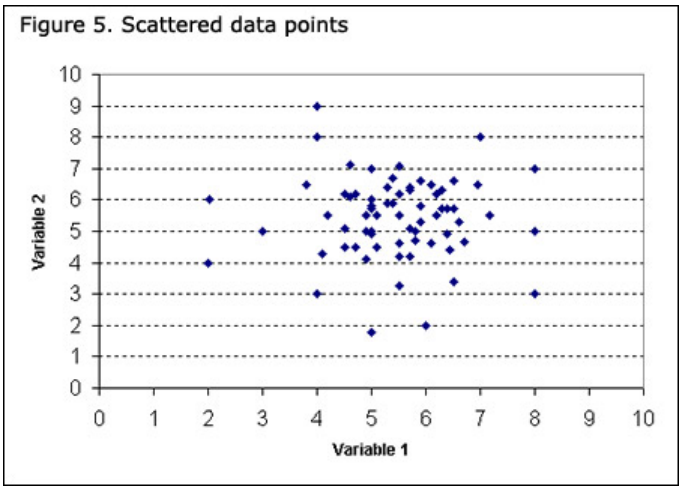
Figure 4. Negative or inverse relationships



Scattered data points

If the data points are randomly scattered, then there is no relationship between the two variables; this means there is a low or zero correlation between the variables (Figure 5).

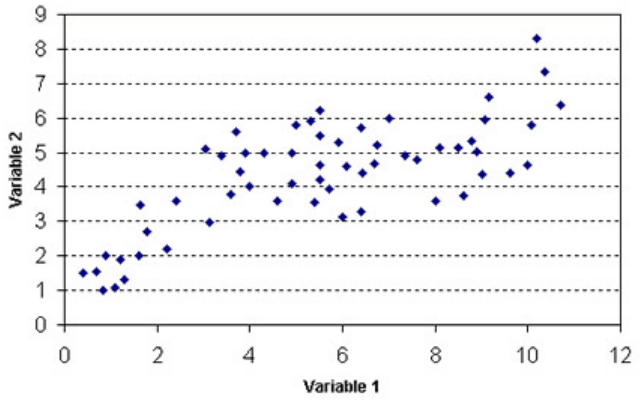
Figure 5. Scattered data points



Non-linear patterns

Very low or zero correlation may result from a non-linear relationship between two variables. If the relationship is, in fact, non-linear (i.e., points clustering around a curve, not a straight line), the linear correlation coefficient will not be a good measure of the strength of the relationship (Figure 6).

Figure 6. Very low or zero correlation



Spread of data

A scatterplot will also illustrate if the data are widely spread or if they are concentrated within a smaller area (Figure 7 and 8).

Figure 7. Data widely spread

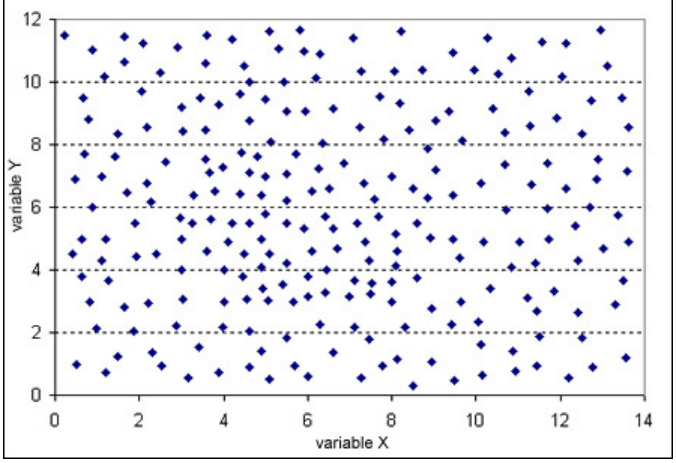
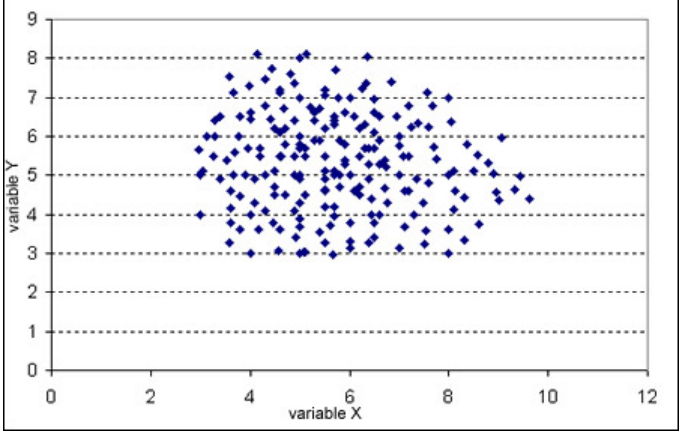


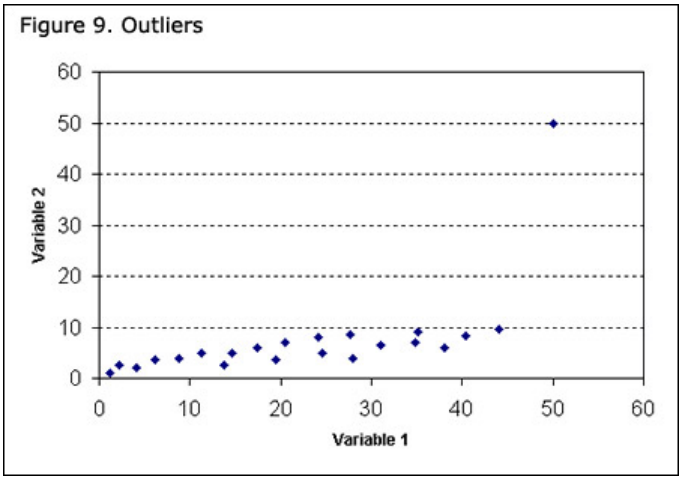
Figure 8. Data concentrated



Outliers

Besides portraying a non-linear relationship between the two variables, a scatterplot can also show whether or not there exist any outliers in the data (Figure 9).

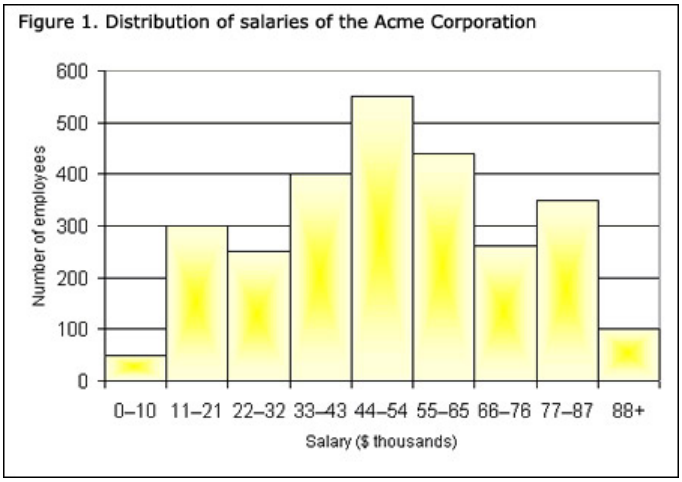
Figure 9. Outliers



Histograms and histographs

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. 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 an area proportional to the number of observations falling into that group. This means that the rectangles will be drawn of non-uniform height. A histogram has an appearance similar to a vertical bar graph, but when the variables are continuous, there are no gaps between the bars. When the variables are discrete, however, gaps should be left between the bars. Figure 1 is a good example of a histogram.

Figure 1. Distribution of salaries of the Acme Corporation



A vertical bar graph and a histogram differ in these ways:

- In a histogram, frequency is measured by the *area* of the column.
- In a vertical bar graph, frequency is measured by the *height* of the bar.

Histogram characteristics

Generally, a histogram will have bars of equal width, although this is not the case when class intervals vary in size. Choosing the appropriate width of the bars for a histogram is very important. As you can see in the example above, the histogram consists simply of a set of vertical bars. Values of the variable being studied are measured on an arithmetic scale along the horizontal x-axis. The bars are of equal width and correspond to the equal class intervals, while the height of each bar corresponds to the frequency of the class it represents.

The histogram is used for variables whose values are numerical and measured on an interval scale. It is generally used when dealing with large data sets (greater than 100 observations). A histogram can also help detect any unusual observations (outliers) or any gaps in the data.

Histographs

A histograph, or frequency polygon, is a graph formed by joining the midpoints of histogram column tops. These graphs are used only when depicting data from the continuous variables shown on a histogram.

A histograph smoothes out the abrupt changes that may appear in a histogram, and is useful for demonstrating continuity of the variable being studied. Figure 2 and 3 are good examples of histographs.

Figure 2. Distribution of salaries for the Acme Corporation

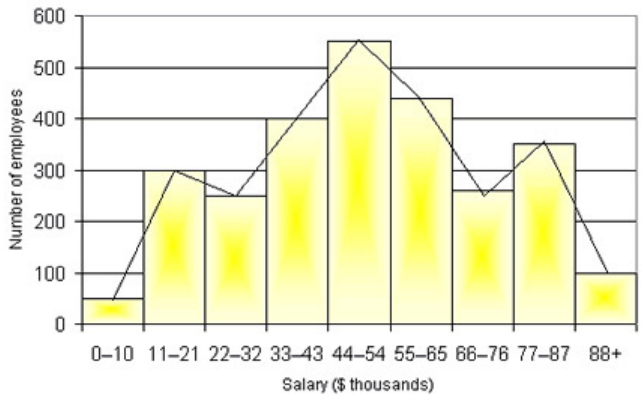
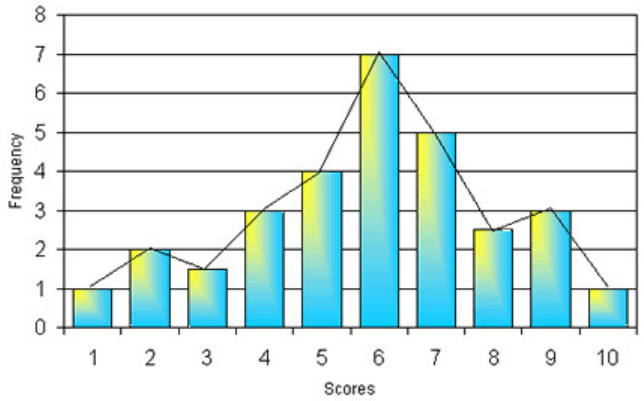


Figure 3. Frequency of scores for a 10-question math quiz



Unlike Figure 2, this histograph has spaces between the bars. By just looking at this illustration, the reader can immediately tell that the spaces mean the variables are discrete. In this way, histographs make it easier for the readers to determine what type of variables has been used.

Summary

If you decide that a graph is the best way to present your information, then no matter what type of graph you use, you need to keep in mind the following 10 tips:

10 tips to make your graphs great!

1. convey an important message
2. decide on a clear purpose
3. draw attention to the message, not the source
4. experiment with various options and graph styles
5. use simple design for complex data
6. make the data 'speak'
7. adapt graph presentation to suit the target audience
8. ensure that the visual perception process is easy and accurate
9. avoid distortion and ambiguity
10. optimize design and integrate style with text and tables

Summary

Graph type	Description
age pyramid	Represents age structure of a population.
vertical bar graph	Compares important data values. Displays data better than horizontal bar graphs, and is preferred when possible.
dot graph	Displays a comparatively large number of categories when category order is unimportant. Best used when portraying category values in descending order.
histogram	Shows discrete or continuous variable data in a similar way to column graphs, but without the gap between the columns.
histograph (frequency polygon)	Depicts continuous variable data. Smoothes abrupt changes which may appear in a histogram
horizontal bar graph	Compares important data. Useful when category names are too long to fit at the foot of a column.
line graph	Often used to depict data over time.
pictograph	Favoured by professional graphic artists, although students can create simple pictorial presentations as well. Comparisons must be accurately depicted and respect the scale.
circle graph/pie chart	Compares a small number of categories. Values should be markedly different, or differences may not be easy to decipher. Labelling pie segments with their actual values overcomes this problem. When data points are similar, the circle graph/pie chart's message may be misunderstood. A bar graph may be better in this case.
scatterplot	Measures two or more variables thought to be related.

Create your own graph

Now you are ready to practice simple graphing on your own. Use the graphing tool at the bottom of this page to help you create your own graph. You simply have to follow the instructions below:

1. Type the title of your graph in the *Graph title* space provided. You can enter up to 80 characters in this space.

   For example, suppose that you wanted to create a bar graph depicting data from a survey on your classmates' favourite school subjects. You might title your graph "Favourite school subjects, by number of students."

2. For a graph with x- and y-axis labels, fill in the labels in the *X-axis heading* and *Y-axis heading* spaces. You can use up to 30 characters in each of the category labels.

   Using the same example, you would label your x-axis with the heading "School subjects" and your y-axis with the heading "Number of students."
   Note: Leave the x-axis heading and x-axis category labels blank for box and whisker plots.

3. In the *X-axis category labels* space, list the various categories that you want to appear on your x-axis or the category slices you want to appear on your circle graph/pie chart. Type in one category per line. Use no punctuation between each category.

   Using the same example, the categories could be "History, Algebra, Biology, English and Drama."

4. Leave the space blank in the *First data set – Legend heading* box, unless you want to compare several types of data on one graph. If you fill in these spaces, then a legend will accompany your graph, showing viewers what each colour represents.

   For example, if you decided to compare the boys' favourite subjects with the girls' favourite subjects, then you would type in "Boys" in the *First data set – Legend heading* box and "Girls" in the *Second data set – Legend heading* box. You may even wish to include a third data set showing the combined figures or total population of "Both boys and girls" favourite subjects.

5. Beneath each *Legend heading* box is another box where you can add the percentages or figures belonging to each category. List each figure corresponding to the category labels in the *X-axis category labels* box.

   Using the same example, beside "History", type in the number "35" for Boys and "27" for Girls. This means that 35 boys and 27 girls said that History was their favourite subject. Continue down the list until all figures are included.

   Note: If you are making a circle graph/pie chart, the numbers will be converted into percentages. If the numbers do not equal 100%, then the numbers will be converted to their equivalent percentage, so that all of the combined slices equal 100%.

6. Select the colour you wish your data set to be from the pull-down list. If you are making a bar graph, then you will want to pick the colour of your bars. If you are making a group bar or line graph, be sure to use different colours for each data set. If you don't change the colours, then by default, the first data set will be in red, the second data set in blue and the third data set in green.

   Note: Circle graph/pie chart colours are automatically set by the program, and box and whisker plots use the same colour for all plots.

7. Click on the boxes that apply to your graph. The *Use origins (0,0)* box is automatically set. However, if you do not want your graph to start from the point of origin and instead want it to start from your first data set, then click on the box to remove the checkmark. If you are making a circle graph/pie chart, then click on the *Show labels on circle graph/pie chart* box. If you want to show each plotted point on your graph, choose the *Show plotted points on graph* box.

8. Select the type of graph you want to create from the pull-down list.

   Note: *Area graphs* resemble line graphs. The points are plotted using a line. However, everything beneath the line is coloured in like bars on a bar graph. This particular graph shows how much area has been covered by the amount of space coloured in. This graph can be used to emphasize trends (sharp declines, flat plateaus, etc.) in the data.

9. Press the *Create graph* button, and your graph should appear on screen. To view your data in different graph types, click the back button and change the graph type from the pull-down list.

Always bear in mind that when using this package that the y values are assumed to be numeric. Anything that is non-numeric is interpreted as 0. The x values are considered labels and therefore will appear as input. Except for box and whisker plots, make sure you enter the label for the X axis (i.e., years, months, days, occupations etc.). If you do not enter the text values in the X-axis your output will not be what you expect.

Graph title:

X-axis heading:

Y-axis heading:

First data set—Legend heading:

Second data set—Legend heading:

Third data set—Legend heading:

X-axis category labels:

First data set Data points:

Second data set Data points:

Third data set Data points:

Pick a colour:

First data set: Light Grey ▼

Second data set: Blue ▼

Third data set: Yellow ▼

- [x] Use origin (0,0)
- [ ] Show plotted points on graph
- [ ] Show labels on circle graph/pie chart

Type of chart: Line graph ▼

Create graph

Exercise

1. Over a period of one week, collect examples of graph types (i.e., bar graphs, circle graphs/pie charts, line graphs) from newspapers, magazines or journals. At the end of the week, look at each of the graph examples and determine whether they are appropriate or whether the statistics would have been better presented with a different type of graph.

2. What type of graph would you use to present the following? Explain your choice.
    a. the number of female students in each grade in your school
    b. the annual number of road fatalities (the road toll) in your province or territory over the last 30 years
    c. the speed (km/h) of the world's 20 fastest animals
    d. the total population of Canada, the provinces and territories

3. Discuss as a class the different situations where a horizontal or a vertical bar graph would be most useful.

Analytical graphing

Numerical variables can be presented in a variety of ways, including *stem and leaf plots* and *frequency distribution tables*, as seen in the section entitled "Organizing data".

Two other methods are presented in this section: *cumulative frequency graphs* and *cumulative percentage graphs*. These analytical graphs are very useful in finding the central tendency of large data sets, as well as providing other valuable information.

Cumulative frequency

Cumulative frequency is used to determine the number of observations that lie above (or below) a particular value in a data set. The cumulative frequency is calculated using a frequency distribution table, which can be constructed from stem and leaf plots or directly from the data.

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.

Discrete or continuous variables

Variables in any calculation can be characterized by the value assigned to them. A *discrete variable* consists of separate, indivisible categories. No values can exist between a variable and its neighbour. For example, if you were to observe a class attendance registered from day-to-day, you may discover that the class has 29 students on one day and 30 students on another. However, it is impossible for student attendance to be between 29 and 30. (There is simply no room to observe any values between these two values, as there is no way of having 29 and a half students.)

Not all variables are characterized as discrete. Some variables (such as time, height and weight) are not limited to a fixed set of indivisible categories. These variables are called *continuous variables*, and they are divisible into an infinite number of possible values. For example, time can be measured in fractional parts of hours, minutes, seconds and milliseconds. So, instead of finishing a race in 11 or 12 minutes, a jockey and his horse can cross the finish line at 11 minutes and 43 seconds.

It is essential to know the difference between the two types of variables in order to properly calculate their cumulative frequency.

Example 1 – Discrete variables

The total rock climber count of 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.

1. Use these discrete variables to:
   - set up a stem and leaf plot, (see the section on stem and leaf plots) with additional columns labelled *Frequency, Upper Value* and *Cumulative frequency*
   - figure out the frequency of observations for each stem
   - find the upper value for each stem
   - calculate the cumulative frequency by adding the numbers in the *Frequency* column
   - record all the results in the plot
2. Plot a graph using the y-axis (or vertical line) for the cumulative frequency and the x-axis (or horizontal line) for the number of people rock climbing.

Answers:

1. The number of rock climbers ranges from 4 to 65. In order to produce a stem and leaf plot, the data are best grouped in class intervals of 10.

   Each interval can be located in the *Stem* column. The numbers within this column represent the first number within the class interval. (For example, *Stem* 0 represents the interval 0–9, *Stem* 1 represents the interval of 10–19, and so forth.)

   The *Leaf* column lists the number of observations that lie within each class interval. For example, in *Stem* 2 (interval 20–29), the three observations, 23, 24, and 26, are represented as 3, 4 and 6.

   The *Frequency* column lists the number of observations found within a class interval. For example, in *Stem* 5, nine leaves (or observations) were found; in *Stem* 1, there are only two.

   Use the *Frequency* column to calculate cumulative frequency.

   - First, add the number from the *Frequency* column to its predecessor. For example, in *Stem* 0, we have only one observation and no predecessors. The cumulative frequency is one.

     1 + 0 = 1

   - However in *Stem* 1, there are two observations. Add these two to the previous cumulative frequency (one), and the result is three.

     1 + 2 = 3

   - In *Stem* 2, there are three observations. Add these three to the previous cumulative frequency (three) and the total (six) is the cumulative frequency for *Stem* 2.

     3 + 3 = 6

- Continue these calculations until you have added up all of the numbers in the *Frequency* column.
- Record the results in the *Cumulative frequency* column.

The *Upper value* column lists the observation (variable) with the highest value in each of the class intervals. For example, in *Stem* 1, the two observations 8 and 9 represent the variables 18 and 19. The upper value of these two variables is 19.

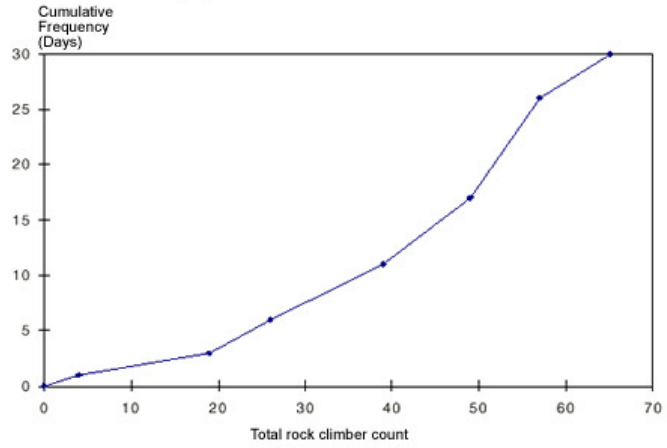
Table 1. Cumulative frequency of daily rock climber counts recorded in Lake Louise, Alberta, 30-day period

Stem	Leaf	Frequency (f)	Upper value	Cumulative frequency
0	4	1	4	1
1	8 9	2	19	1 + 2 = 3
2	3 4 6	3	26	3 + 3 = 6
3	1 5 5 7 9	5	39	6 + 5 = 11
4	0 1 2 3 5 9	6	49	11 + 6 = 17
5	0 1 1 2 4 4 5 6 7	9	57	17 + 9 = 26
6	0 2 3 5	4	65	26 + 4 = 30

2. Since these variables are discrete, use the upper values in plotting the graph. Plot the points to form a continuous curve called an ogive.

   Always label the graph with the cumulative frequency—corresponding to the number of observations made—on the vertical axis. Label the horizontal axis with the other variable (in this case, the total rock climber counts) as shown below:

Figure 1. Number of rock climbers in Lake Louise, Alberta, 30-day period



The following information can be gained from either graph or table:

- on 11 of the 30 days, 39 people or fewer climbed the rocks around Lake Louise
- on 13 of the 30 days, 50 or more people climbed the rocks around Lake Louise

When a continuous variable is used, both calculating the cumulative frequency and plotting the graph require a slightly different approach from that used for a discrete variable.

Example 2 – Continuous variables

For 25 days, the snow depth at Whistler Mountain, B.C. was measured (to the nearest centimetre) and recorded as follows:

242, 228, 217, 209, 253, 239, 266, 242, 251, 240, 223, 219, 246, 260, 258, 225, 234, 230, 249, 245, 254, 243, 235, 231, 257.

1. Use the continuous variables above to:
   - set up a frequency distribution table
   - find the frequency for each class interval
   - locate the endpoint for each class interval
   - calculate the cumulative frequency by adding the numbers in the *Frequency* column
   - record all results in the table
2. Use the information gathered from the frequency distribution table to plot a cumulative frequency graph.

Answers:

1. The snow depth measurements range from 209 cm to 266 cm. In order to produce the frequency distribution table, the data are best grouped in class intervals of 10 cm each.

   In the *Snow depth* column, each 10-cm class interval from 200 cm to 270 cm is listed.

   The *Frequency* column records the number of observations that fall within a particular interval. This column represents the observations in the *Tally* column, only in numerical form.

The *Endpoint* column functions much like the *Upper value* column of Exercise 1, with the exception that the endpoint is the highest number in the interval, regardless of the actual value of each observation. For example, in the class interval of 210–220, the actual value of the two observations is 217 and 219. But, instead of using 219, the endpoint of 220 is used.

The *Cumulative frequency* column lists the total of each frequency added to its predecessor.

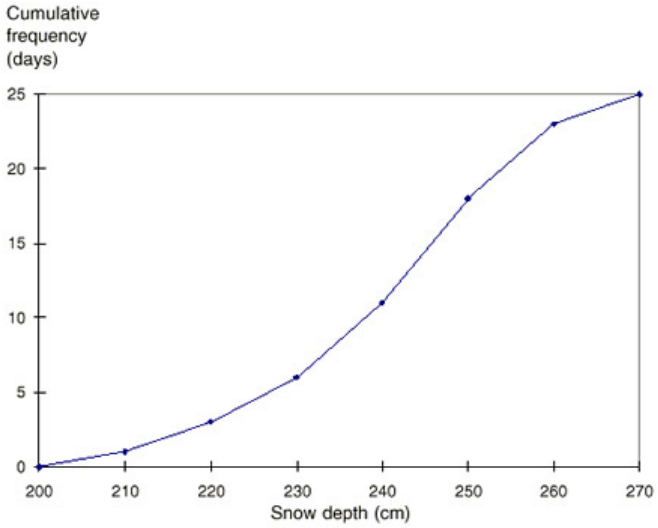
Table 2. Snow depth measured at Whistler Mountain, B.C., 25-day period

Snow depth (*x*)	Tally	Frequency (f)	Endpoint	Cumulative frequency
			200	0
200 to < 210	I	1	210	1
210 to < 220	II	2	220	3
220 to < 230	III	3	230	6
230 to < 240	𝍸	5	240	11
240 to < 250	𝍸 II	7	250	18
250 to < 260	𝍸	5	260	23
260 to < 270	II	2	270	25

2. Because the variable is continuous, the endpoints of each class interval are used in plotting the graph. The plotted points are joined to form an ogive.

   Remember, the cumulative frequency (number of observations made) is labelled on the vertical y-axis and any other variable (snow depth) is labelled on the horizontal x-axis as shown in Figure 2.

Figure 2. Snow depth measured at Whistler Mountain, B.C., 25-day period



The following information can be gained from either graph or table:

- none of the 25 days had snow depth less than 200 cm
- one of the 25 days snow had depth of less than 210 cm
- two of the 25 days snow had depth 260 cm or more

Other cumulative frequency calculations

Another calculation that can be obtained using a frequency distribution table is the *relative frequency distribution*. This method is defined as the percentage of observations falling in each class interval. Relative cumulative frequency can be found by dividing the frequency of each interval by the total number of observations. (For more information, see Frequency distribution in the chapter entitled Organizing data.)

A frequency distribution table can also be used to calculate *cumulative percentage*. This method of frequency distribution gives us the percentage of the cumulative frequency, as opposed to the percentage of just the frequency.

Cumulative percentage

Cumulative percentage is another way of expressing frequency distribution. It calculates the percentage of the cumulative frequency within each interval, much as relative frequency distribution calculates the percentage of frequency.

The main advantage of cumulative percentage over cumulative frequency as a measure of frequency distribution is that it provides an easier way to compare different sets of data.

Cumulative frequency and cumulative percentage graphs are exactly the same, with the exception of the vertical axis scale. In fact, it is possible to have the two vertical axes, (one for cumulative frequency and another for cumulative percentage), on the same graph.

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 percentage = (cumulative frequency ÷ n) x 100

Example 1 – Calculating cumulative percentage

For 25 days, the snow depth at Whistler Mountain, B.C. was measured (to the nearest centimetre) and recorded as follows:

242, 228, 217, 209, 253, 239, 266, 242, 251, 240, 223, 219, 246, 260, 258, 225, 234, 230, 249, 245, 254, 243, 235, 231, 257.

1. Use the data above (the same data as in Example 2 of the previous section on Cumulative Frequency) to:
   - construct another frequency distribution table
   - figure out what the frequency is for each interval
   - find out the endpoint for each interval
   - calculate the cumulative frequency and percentage
   - record results in the table
2. Draw a graph with two different vertical (y) axes (either on each side of the graph, or side by side): one for cumulative frequency and one for cumulative percentage. Be sure to label *cumulative frequency* and *cumulative percentage* on either side of the vertical or y-axis. Label the x-axis with the other variable (snow depth).

Answers:

1. The snow depth measurements range from 209 cm to 266 cm. In order to produce the table, the data are best grouped in class intervals of 10 cm each.

   In the *Snow depth* column, each 10-cm class interval from 200 cm to 270 cm is listed.

   The *Frequency* column records the number of observations that fall within a particular interval. This column represents the observations in the *Tally* column, only in numerical form.

   Each of the numbers in the *Endpoint* column is the highest number in each class interval. In the interval of 200 cm to 210 cm, the endpoint would be 210.

   The *Cumulative frequency* column lists the total of each frequency added to its predecessor, as seen in the exercises in the previous section.

   The *Cumulative percentage* column divides the cumulative frequency by the total number of observations (in this case, 25). The result is then multiplied by 100. This calculation gives the cumulative percentage for each interval.

Table 1. Snow depth measured at Whistler Mountain, B.C., 25-day period

Snow depth (*x*)	Tally	Frequency (f)	Endpoint	Cumulative frequency	Cumulative percentage
			200	0	0 ÷ 25 x 100 = 0
200 to < 210	I	1	210	1	1 ÷ 25 x 100 = 4
210 to < 220	II	2	220	3	3 ÷ 25 x 100 = 12
220 to < 230	III	3	230	6	6 ÷ 25 x 100 = 24
230 to < 240	𝍸	5	240	11	11 ÷ 25 x 100 = 44
240 to < 250	𝍸 II	7	250	18	18 ÷ 25 x 100 = 72
250 to < 260	𝍸	5	260	23	23 ÷ 25 x 100 = 92
260 to < 270	II	2	270	25	25 ÷ 25 x 100 = 100

2. Apart from the extra axis representing the cumulative percentage, the graph should look exactly the same as that drawn in Example 2 of the section on Cumulative frequency.

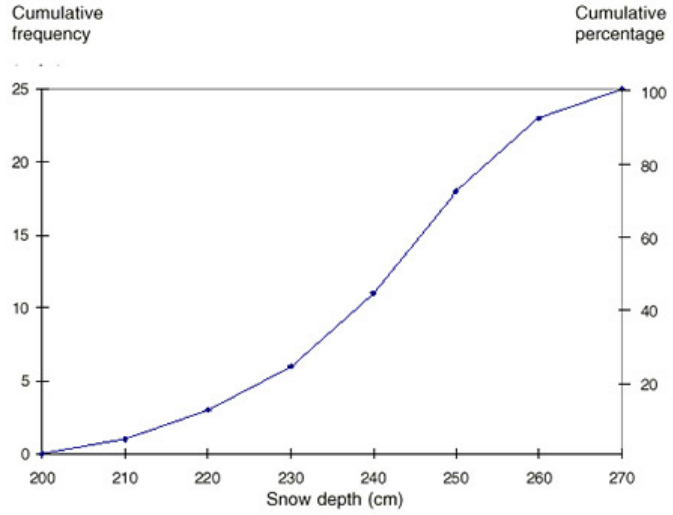
   The *Cumulative percentage* axis is divided into five intervals of 20, while the *Cumulative frequency* axis is divided into five intervals of 5. The *Snow depth* axis is divided by the endpoints of each 10-cm class interval.

Using each endpoint to plot the graph, you will discover that both the cumulative frequency and the cumulative percentage land in the same spot. For example, using the endpoint of 260, plot your point on the 23rd day (cumulative frequency). This point happens to be in the same place where the cumulative percentage (92%) will be plotted.

You have to be very careful when you are building a graph with two y-axes. For example, if you have 47 observations, you might be tempted to use intervals of 5 and end your y-axis at the cumulative frequency of 50. However, when you draw your y-axis for the cumulative percentage, you must put the 100% interval at the same level as the 47 mark on the other y-axis—not at the 50 mark. For this example, a cumulative frequency of 47 represents 100% of your data. If you put the 100% at the top of the scale where the 50 interval is marked, your line for the cumulative frequency will not match the line for the cumulative percentage.

The plotted points join to form an ogive, which often looks similar to a stretched ***S***. Ogives are used to determine the number, or percentage, of observations that lie above or below a specified value. For example, according to the table and the graph, 92% of the time the snow depth recorded in the 25-day period was below the 260 cm mark.

Figure 1. Snow depth measured at Whistler Mountain, B.C., 25-day period



The following information can be gained from either the graph or table:

- during the 25-day period, 24% of the time the recorded snow depth was less than 230 cm
- on 7 of the 25 days, snow depth was at least 250 cm

Exercices

1. The following set of data gives the length of reign in years of the various Kings and Queens of England since the Battle of Hastings in 1066. If the monarch's reign lasted less than six months, then the number was rounded down to zero. Otherwise, the number was rounded to the nearest year.

Monarch

Monarch	Years	Monarch	Years
William the Conquerer	24	Edward VI	7
William II	13	Jane	0
Henry I	34	Mary I	5
Stephen	18	Elizabeth I	44
Matilda	1	James I	22
Henry II	34	Charles I	24
Richard I	10	Charles II	25
John	18	James II	3
Henry III	56	William III	13
Edward I	35	Mary II	5
Edward II	20	Anne	12
Edward III	50	George I	13
Richard II	22	George II	33
Henry IV	14	George III	59
Henry V	9	George IV	10
Henry VI	39	William IV	7
Edward IV	22	Victoria	64
Edward V	0	Edward VII	9
Richard III	2	George V	26
Henry VII	23	Edward VII	1
Henry VIII	38	George VI	15

   a. Present the data in the form of an ordered stem and leaf plot.
   b. Do any outliers exist? If so, give a reason for their presence.
   c. Describe the following features of distribution:
      - the number of peaks
      - the general shape
      - the approximate value of the centre of the distribution

   d. Calculate cumulative frequency and cumulative percentage.
   e. Draw the ogive with two different vertical axes, one for cumulative frequency and one for cumulative percentage.
   f. Using the information found earlier in this exercise, answer the following questions:
      - How many monarchs have reigned for less than 10 years?
      - How many monarchs have reigned for 50 years or more?

   g. The current English monarch is Queen Elizabeth II. She has ruled since February 6, 1952, and has been excluded from the original data set. Calculate the length of her reign, and briefly comment on it in comparison with the other rulers.

2. Eliza often buys a small bag of french fries at *Hungry Stats*, the local fast-food stand. She was curious to know whether or not she was getting value for her money and was also interested in finding out whether or not she received the same amount of french fries in each bag. So, for three months, Eliza decided to count and record the fries in each bag. At the end of the third month, she realized that she had bought fries a total of 30 times. The resulting data from her 30 visits are as follows:

   44, 46, 54, 38, 49, 46, 45, 31, 55, 37, 42, 43, 47, 51, 48 40, 59, 35, 47, 21,43, 37, 45, 38, 40, 32, 50, 34, 43, 54

   a. Present the data in an ordered stem and leaf plot. If necessary, split the stems.
   b. Do any outliers exist? If so, give a reason for their presence?
   c. Describe the following features of distribution:
      - the number of peaks

- the general shape
- the approximate value of the centre of the distribution

d. Calculate cumulative frequency and cumulative percentage.

e. Draw the ogive with two different vertical axes, one for cumulative frequency and one for cumulative percentage.

f. Using the information found earlier in this exercise, answer the following questions:

- How many bags had fewer than 40 fries?
- What percentage of bags had 45 or more fries?

g. *Hungry Stats* decided to create a slogan for their new advertising campaign. Complete the following statement using the information gained from Eliza's calculations.

"Our bags may be small, but half contain at least___ french fries!"

3. Imagine that the following table represents (by age group) the number of unemplyed female job-seekers for full-time work.

Table 1. Number of unemployed female job-seekers, by age group

Age group *	Number of females
15 to 24	339
25 to 34	273
35 to 44	147
45 to 54	121
55 to 64	22

* Age is collected in completed number of years. Thus, the interval 15 to 24 has an upper endpoint of 25 (refer to Stem and leaf plots in the chapter on Organizing data).

a. Are the age variables discrete or continuous?

b. Copy Table 1 into your notebook and calculate the cumulative frequency and cumulative percentage.

c. Draw the ogive with two different vertical axes, one for cumulative frequency and one for cumulative percentage.

d. Why are there no data for females younger than 15 years old?

e. In what age group does the cumulative percentage value 50% lie?

f. What percentage of unemployed female job-seekers are younger than 25 years old?

g. What percentage of unemployed female job-seekers are 55 years old or older?

h. How would the Canadian government use this sort of information?

4. Imagine a survey was conducted in order to determine how long it takes for Statistics Canada employees to travel to work. The results, to the nearest minute, were recorded as follows:

33, 63, 49, 65, 56, 45, 52, 63, 38, 66, 43, 98, 60, 58, 68, 29, 59, 87, 22, 64, 73, 56, 71, 67, 44, 31, 83, 50, 75, 65, 60, 51, 89, 69, 41, 76,58, 62, 25, 52, 64, 77, 61, 55, 80, 45, 12, 69, 40, 37

a. Are these variables discrete or continuous?

b. Present the data in a frequency table, using appropriate intervals; including relative and percentage frequencies.

c. Draw a histogram to represent the data and plot the frequency polygon.

d. Prepare an ordered stem and leaf plot for the data. Do any outliers exist? If so, give a reason for their presence.

e. Describe the following features distribution:

- the number of peaks
- the general shape
- the approximate value of the centre of the distribution

f. Calculate the cumulative frequency and cumulative percentage. Find the endpoints and record the results in a table.

g. Draw the ogive with two different vertical axes, one for cumulative frequency and one for cumulative percentage.

h. Using the information found earlier in the exercise, answer the following questions:

- With regards to how long it takes to travel to work, what was the most common time interval for Statistics Canada employees?
- What percentage of employees took longer than 90 minutes to travel to work?
- How many employees took less than 40 minutes to travel to work?

Class activity

Survey the teachers in your school to find out how long they have been teaching (to the nearest year).

1. Are the variables discrete or continuous?

2. Present the data in a frequency table, using appropriate intervals, including relative and percentage frequencies.
3. Using the information found earlier in this exercise, answer the following questions:
    a. How many years have the majority of teachers been teaching?
    b. What is the percentage difference between the first and the second most common length of service?
4. Draw a histogram to represent the data and plot the frequency polygon. Prepare an ordered stem and leaf plot for the data.
5. Do any outliers exist? If so, give a reason for their presence.
6. Describe the following features of distribution:
    - the number of peaks
    - the general shape
    - the approximate value of the centre of the distribution
7. Calculate cumulative frequency and cumulative percentage.
8. Draw the ogive with two different vertical axes, one for cumulative frequency and one for cumulative percentage.
9. With the new information, answer the following questions:
    a. How many teachers have taught for more than 10 years?
    b. What percentage of teachers have taught for more than 10 years?
    c. What percentage of teachers have taught for less than 10 years?
    d. What is the number of years below which half the teachers have taught?
10. Present your analysis in a report containing the necessary cumulative frequency and percentage tables and graphs.

Answers

1.

a.

Table 1. Number of years of reign for English monarchs

Stem	Leaf	Frequency (f)	Upper Value	Cumulative frequency	Cumulative percentage
0	0 0 1 1 2 3 5 5 7 7 9 9	12	9	12	28.57
1	0 0 2 3 3 3 4 5 8 8	10	18	22	52.23
2	0 2 2 2 3 4 4 5 6	9	26	31	73.80
3	3 4 4 5 8 9	6	38	37	88.09
4	4	1	44	38	90.47
5	0 6 9	3	59	41	97.61
6	4	1	64	42	100.00

b. The possible outliers are 56, 59 and 64. However, this data is based on fact, so the numbers exist because each of these monarchs came to the throne early in their youth and enjoyed long lives. This reasoning suggests that there are no outliers.

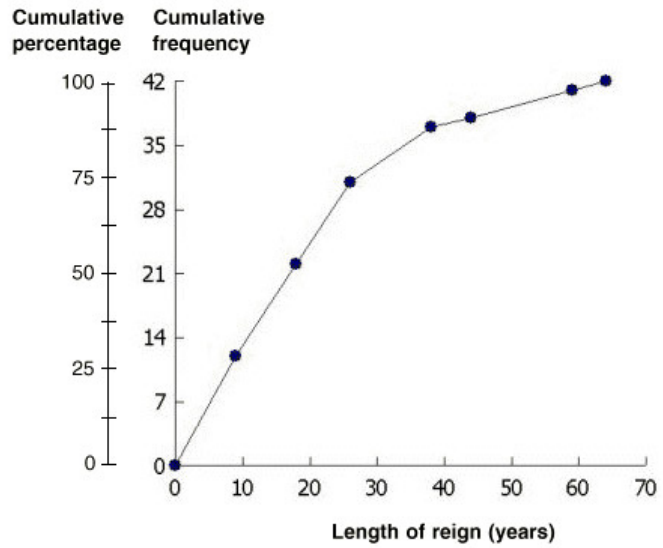
c.

- The number of peaks that appear at the beginning of the distribution is one.
- The general shape of the distribution is skewed to the right.
- The approximate value at the centre of the distribution is 18 years.

d. Same answer of table 1

e.

Figure 1. Number of years of reign for English monarchs



f.

- There are 12 monarchs who have reigned for less than 10 years.
- There are 4 monarchs who have reigned for 50 years or more.

g. Queen Elizabeth II celebrated her 58th year of rulership on February 6, 2010. Her reign is already well above the centre of distribution and only two other monarchs have been on the throne longer.

2.

a.

Table 2. Number of french fries per small bag

Stem	Leaf
2(0)	1
2(5)	
3(0)	1 2 4
3(5)	5 7 7 8 8
4(0)	0 0 2 3 3 3 4
4(5)	5 5 6 6 7 7 8 9
5(0)	0 1 4 4
5(5)	

b. The outlier in this exercise is 21. One reason could be that there were only 21 fries in the bag on that day because they were the only fries left in the batch. Another reason could be that the number recorded was incorrect (i.e., 21 instead of 41).

c.

- The graph is unimodal, meaning it has only one peak in this distribution.
- If the outlier is removed, the general shape of the distribution is roughly symmetric.
- The approximate value at the center of distribution is between 43 and 44 or 43.5 fries.

d.

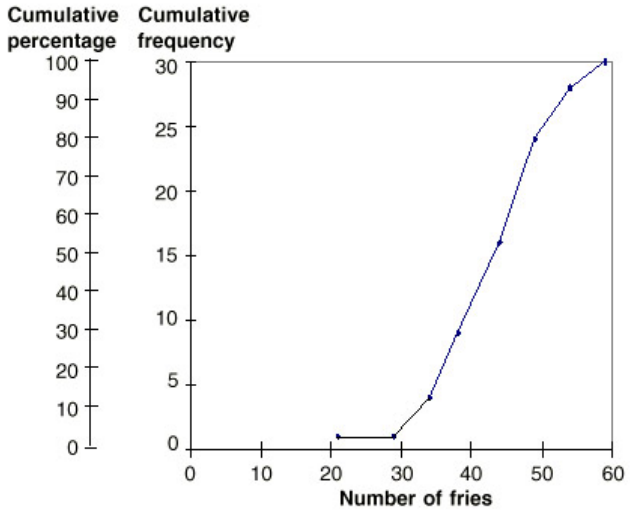
Table 3. Number of french fries per small bag

Number of fries	Frequency (f)	Upper value	Cumulative frequency	Cumulative percentage
20 to 24	1	21	1	3.3
25 to 29	0	29	1	3.3
30 to 34	3	34	4	13.3
35 to 39	5	38	9	30.0
40 to 44	7	44	16	53.3
45 to 49	8	49	24	80.0
50 to 54	4	54	28	93.3
55 to 59	2	59	30	100.0
Total	**30**			**100.0**

Note: If you have a class interval that is empty, you should always use the endpoint as the upper value. For instance, in the above example, there is one bag in the 20–24 interval, but no bags in the 25–29 interval. To determine the upper value for the 25–29 interval, use the endpoint of 29.

e.

Figure 2. Number of french fries per small bag



f.

- In 30 bags of french fries, only 9 had fewer than 40 fries in them.
- The percentage of bags with 45 or more fries is 46.7%.

g. The promotional slogan should read: "Our bags may be small but half contain at least 44 french fries!"

3.

a. These are continuous variables.

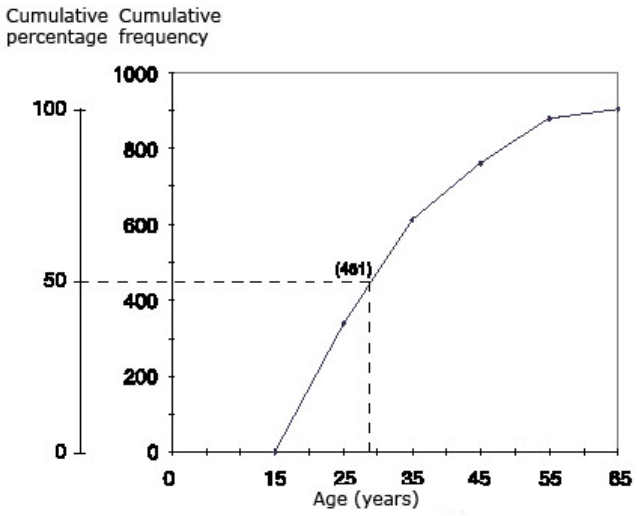
b.

Table 4. Number of unemployed female job-seekers, by age group

Age group	Number of females	Endpoint	Cumulative frequency	Cumulative percentage
0 to 14	0	15	0	0.0
15 to 24	339	25	339	37.6
25 to 34	273	35	612	67.8
35 to 44	147	45	759	84.1
45 to 54	121	55	880	97.6
55 to 64	22	65	902	100.0

c.

Figure 3. Number of unemployed female job seekers, by age group



d. There is no data for females under 15 years of age because no one under 15 can be classified as unemployed.

e. The cumulative percentage of 50% falls within the age group of 25–34 (approximately 29 years old).

f. The percentage of unemployed females who are under 25 years of age and looking for full-time work is 37.6%.

g. The percentage of unemployed females who are 55 years and older and looking for full-time work is 2.4%.

h. The Canadian government could establish job-creation schemes directed at particular age groups. In this case, the job-creation scheme would likely be for those under 25 years of age.

4.

a. These are continuous variables.

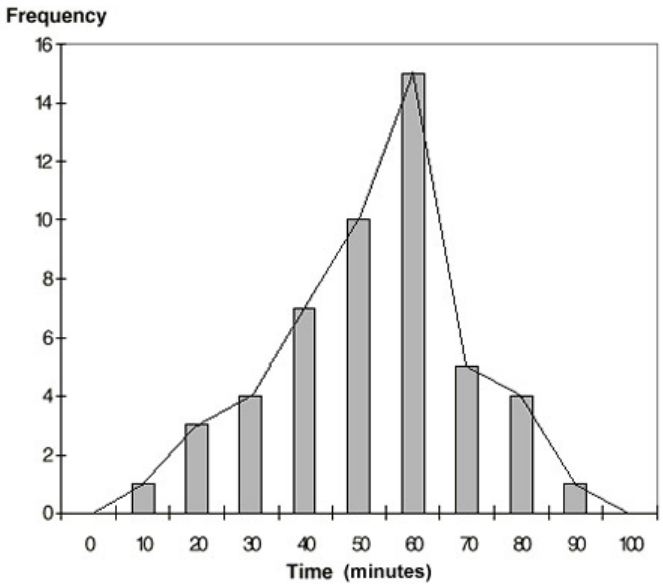
b.

Table 5. Commuter time of Statistics Canada employees, Ottawa

Time (x)	Tally	Frequency (f)	Relative frequency	Relative percentage
0 to < 10		0	0.00	0
10 to < 20	I	1	0.02	2
20 to < 30	III	3	0.06	6
30 to < 40	IIII	4	0.08	8
40 to < 50	𝍸 II	7	0.14	14
50 to < 60	𝍸 𝍸	10	0.20	20
60 to < 70	𝍸 𝍸 𝍸	15	0.30	30
70 to < 80	𝍸	5	0.10	10
80 to < 90	IIII	4	0.08	8
90 to < 100	I	1	0.02	2
Total		**50**	**1.00**	**100**

c.

Figure 4. Commuter time for Statistics Canada employees, Ottawa



d.

Table 6. Commuter time of Statistics Canada employees, Ottawa

Stem	Leaf
0	
1	2
2	2 5 9
3	1 3 7 8
4	0 1 3 4 5 5 9
5	0 1 2 2 5 6 6 8 8 9
6	0 0 1 2 3 3 4 4 5 5 6 7 8 9 9
7	1 3 5 6 7
8	0 3 7 9
9	8

A possible outlier could be 98. The reason for this outlier might be that the person had difficulty in getting to work, or simply lives further away than most employees.

e.

- The graph is unimodal, meaning it only has one peak in the distribution.
- The general shape at the centre of distribution is quite symmetric.
- The approximate value at the centre of distribution is between 59 and 60 or 59.5 minutes.

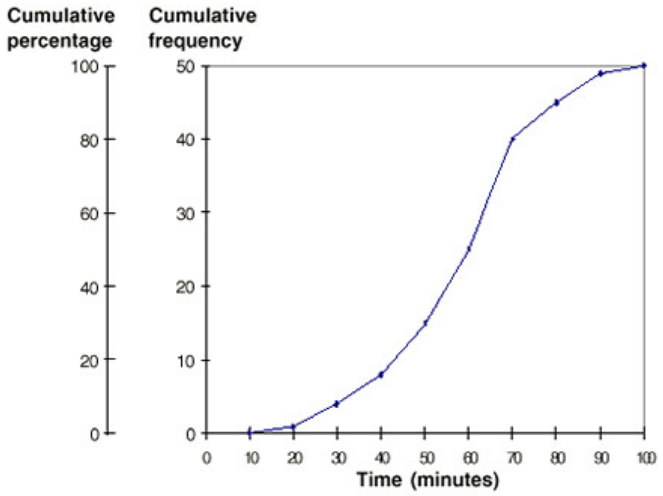
f.

Table 7. Commuter time of Statistics Canada employees, Ottawa

Time (x)	Frequency (f)	Endpoint	Cumulative frequency	Cumulative percentage
0 to < 10	0	10	0	0
10 to < 20	1	20	1	2
20 to < 30	3	30	4	8
30 to < 40	4	40	8	16
40 to < 50	7	50	15	30
50 to < 60	10	60	25	50
60 to < 70	15	70	40	80
70 to < 80	5	80	45	90
80 to < 90	4	90	49	98
90 to < 100	1	100	50	100

g.

Figure 5. Commuter time for Statistics Canada employees, Ottawa



h.

- The most common time interval for Statistics Canada employees to get to work in is 60 –< 70 minutes.
- Only 2% of employees take more than 90 minutes to travel to work.
- Out of the 50 employees surveyed, only 8 took less than 40 minutes to travel to work.

Measures of central tendency

The best way to reduce a set of data and still retain its information is to summarize it with a single value.

Measures of *central tendency*—mean, median, and mode—can help you capture, with a single number, what is typical of a data set.

The *mean* is the average value of all the data in the set.

The *median* is the middle value in a data set that has been arranged in numerical order so that exactly half the data is above the median and half is below it.

The *mode* is the value that occurs most frequently in the set.

In a *normal distribution*, mean, median and mode are identical in value.

Calculating the mean

The mean of a numeric variable is calculated by adding the values of all observations in a data set and then dividing that sum by the number of observations in the set. This provides the average value of all the data.

Mean = sum of all the observation values ÷ number of observations

There are two types of variables—*discrete* and *continuous*. Discrete variables are defined as variables that cannot be divided internally. For example, a hockey player can score 1 or 2 goals, but never 1 and a half goals. Continuous variables, however, can be divided into smaller units. A student's age can be 11 years, 7 months and 3 days, as opposed to just 11 or 12 years.

It is important that you understand the difference between these two types of variables, so that you can properly calculate the mean in any given situation. The following examples use discrete variables to calculate the mean

Example 1 – Soccer tournament at Mount Rival I

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 was the mean score?

Mean = sum of all the observed values ÷ number of observations

= (7 + 5 + 0 + 7 + 8 + 5 + 5 + 4 + 5 + 1) ÷ 10
= 47 ÷ 10
= 4,7

Therefore, in the 10-game tournament, the player scored an average of 4.7 goals per game. The average of 4.7 is not a whole number so it only has meaning in a statistical sense. In reality, it is impossible to score 4.7 goals, even if you are a top scorer.

The mathematical notation to calculate the mean for a discrete variable is as follows:

$$\overline{X} = \frac{\Sigma x}{n}$$

ou

Σ
x ÷ n

where **x** stands for an observed value,

n stands for the number of observations in the data set,

Σ

x stands for the sum of all observed **x** values, and

$\bar{x}$
stands for the mean value of **x**.

Example 2 – Traffic fatalities

The following table lists the number of people killed in traffic accidents over a 10 year period. During this time period, what was the average number of people killed per year? How many people died each day on average in traffic accidents during this time period?

Table 1. Number of fatalities in traffic accidents

Year	Fatalities
1	959
2	1,037
3	960
4	797
5	663
6	652
7	560
8	619

9	623
100	583

Using the formula to calculate the mean for discrete variables, you can see that:

Mean =

Σ

x ÷ n

= (959 + 1 037 + 960 + 797 + 663 + 652 + 560 + 619 + 623 + 583) ÷ 10
= 7,453 ÷ 10
= 745.3

The average number of people killed per year is 745.3.

To calculate the daily death rate from traffic accidents, the average yearly death rate is divided by the number of days in a year (leap years are ignored).

= 745.3 ÷ 365
= 2.0

Therefore, on average, 2 people died each day in traffic accidents.

Frequency tables

A frequency table lists the number of observations that lie in any given data set. It can be used with grouped or ungrouped variables.

For example, to provide a frequency table of the age of people in a data set, you can produce a table using the exact age (ungrouped), or you can group the ages (grouped).

An ungrouped variable can be regarded as being a special type of grouped variable (i.e. (that is), a group). You can calculate the mean of a discrete variable using a frequency table. This method provides an approximation of the true mean for an ungrouped variable. How accurate the approximation is depends on how evenly the observed values are spread within each group.

Example 3 – Soccer tournament at Mount Rival II

Grouping observations in tables is useful when dealing with a large amounts of data. The goal-scoring figures from the soccer tournament example can be displayed in a frequency table.

Table 2. Mount Rival soccer tournament, frequency of goals for lead scorer

Number of goals (x)	Frequency (f)	Total number of goals (xf)
0	1	0
1	1	1
4	1	4
5	4	20
7	2	14
8	1	8
Total (Σ)	10	47

Because the observations are grouped, the mathematical notation changes slightly.

For a discrete variable in a frequency table, the mean is calculated as follows:

$\bar{x} = \frac{\Sigma xf}{\Sigma f}$

ou

Σ xf ÷ Σ f

- where **x** stands for an observed value,
- **xf** stands for the product of an observed value, multiplied by its frequency,
- Σ **xf** stands for the total of all **xf** values,
- Σ **f** stands for the total of all frequencies, and
- $\bar{x}$ stands for the mean value of **x**.

The calculation for the mean of the player's goals is:

Mean = Σ xf ÷ Σ f

= (0 + 1 + 4 + 20 + 14 + 8) ÷ (1 + 1 + 1 + 4 + 2 + 1)
= 47 ÷ 10
= 4,7

Since the variable is ungrouped, this is the exact mean. The next example shows what happens when working with grouped variables.

Calculating the median

If observations of a variable are ordered by value, the median value corresponds to the middle observation in that ordered list. The median value corresponds to a cumulative percentage of 50% (i.e., 50% of the values are below the median and 50% of the values are above the median). The position of the median is

{(n + 1) ÷ 2}th value, where **n** is the number of values in a set of data.

In order to calculate the median, the data must first be ranked (sorted in ascending order). The median is the number in the middle.

Median = the middle value of a set of ordered data.

The median is usually calculated for numeric variables, but may also be calculated for categorical variables that are sequenced, such as the categories in a satisfaction survey: excellent, good, satisfactory and poor. These qualitative categories can be ranked in order, and thus, are considered ordinal.

Raw data

In raw data, the median is the point at which exactly half of the data are above and half below. These halves meet at the median position. If the number of observations is odd, the median fits perfectly and the depth of the median position will be a whole number. If the number of observations is even, the depth of the median position will include a decimal. You need to find the midpoint between the numbers on either side of the median position.

Example 1 – Raw data (discrete variables)

Imagine that a top running athlete in a typical 200-metre training session runs in the following times:

26.1, 25.6, 25.7, 25.2 et 25.0 seconds.

How would you calculate his median time?

First, the values are put in ascending order: 25.0, 25.2, 25.6, 25.7, 26.1. Then, using the following formula, figure out which value is the middle value. Remember that n represents the number of values in the data set.

Median = {(n + 1) ÷ 2}th value
= (5 + 1) ÷ 2
= 3

The third value in the data set will be the median. Since 25.6 is the third value, 25.6 seconds would be the median time.

= 25.6 secondes

Example 2 – Raw data (discrete variables)

Now, if the runner sprints the sixth 200-metre race in 24.7 seconds, what is the median value now?

Again, you first put the data in ascending order: 24.7, 25.0, 25.2, 25.6, 25.7, 26.1. Then, you use the same formula to calculate the median time.

Median = {(n + 1) ÷ 2}th value
= (6 + 1) ÷ 2
= 7 ÷ 2
= 3,5

Since there is an even number of observations in this data set, there is no longer a distinct middle value. The median is the 3.5th value in the data set meaning that it lies between the third and fourth values. Thus, the median is calculated by averaging the two middle values of 25.2 and 25.6. Use the formula below to get the average value.

Average = (value below median + value above median) ÷ 2
= (third value + fourth value) ÷ 2
= (25.2 + 25.6) ÷ 2
= 50.8 ÷ 2
= 25.4

The value 25.4 falls directly between the third and fourth values in this data set, so 25.4 seconds would be the median time.

Ungrouped frequency distribution

In order to find the median using cumulative frequencies (or the number of observations that lie above or below a particular value in a data set), you must calculate the first value with a cumulative frequency greater than or equal to the median. If the median's value is exactly 0.5 more than the cumulative frequency of the previous value, then the median is the midpoint between the two values.

Example 3 – Ungrouped frequency table (discrete variables)

Imagine that your school baseball team scores the following number of home runs in 10 games:

4, 5, 8, 5, 7, 8, 9, 8, 8, 7

If you were to place the total home runs in a frequency table, what would the median be?

First, put the scores in ascending order:

4, 5, 5, 7, 7, 8, 8, 8, 8, 9

Then, make a table with two columns. Label the first column "Number of home runs" and then list the possible number of home runs the team could get. You can start from 0 and list up until the number 10, but since the team never scored less than 4 home runs, you may wish to start listing at the number 4.

Label the second column "Frequency." In this column, record the number of times 4 home runs were scored, 5 home runs were scored and so on. In this case, there was only one time that 4 home runs were scored, but two times that 5 home runs were scored. If you add all of the numbers in the Frequency column, they should equal 10 (for the 10 games played).

Table 1. Number of home runs in 10 baseball games

Number of home runs (x)	Frequency (f)
4	1
5	2
6	0
7	2
8	4
9	1

To find the median, again use the same formula:

Median = {(n + 1) ÷ 2}th value
= (10 + 1) ÷ 2
= 11 ÷ 2
= 5.5
= the median is the 5.5th value in the data set

To get the median, add up the numbers in the Frequency column until you get to 5 (and since the total number of games is 10, the remaining numbers in that column should also equal 5). You will reach 5 after adding all of the frequencies up to and including those for the 7 home runs. The next set of five will begin with the frequencies for 8 home runs. The median (the 5.5th value) lies between the fifth value and the sixth value. Thus, the median lies between 7 home runs and 8 home runs.

If you calculate the average of these values (using the same formula used in Example 2), the result is 7.5.

Average = (middle value before + middle value after) ÷ 2
= (fifth value + sixth value) ÷ 2
= (7 + 8) ÷ 2
= 15 ÷ 2
= 7.5

Technically, the median should be a possible variable. In the above example, the variables are discrete and always whole numbers. Therefore, 7.5 is not a possible variable—no one can hit 7 and a half home runs. Thus, this number only makes sense statistically. Some mathematicians may argue that 8 is a more appropriate median.

Grouped frequency distribution

Sometimes it does not make sense to list each individual variable when a frequency distribution table would be long and cumbersome to work with. In order to simplify this, divide the range of data into intervals and then list the intervals in a frequency distribution table, including a column for the cumulative percentage. (For more information, refer to the Cumulative frequency section.)

The calculation to find the median is a little longer because the data have been grouped into intervals and, therefore, all of the original information has been lost. Some textbooks simply take the midpoint of the interval as the median. However, that method is an over-simplification of the true value. Use the following calculations to find the median for a grouped frequency distribution.

1. Figure out which interval contains the median by using the **(n + 1) ÷ 2** formula. Take whatever value the calculation gives you and then add up the numbers in the frequency column until you come to that value (just like Example 3). For example, if your median is the 13.5th value, add up the frequencies until you come to the 13th and 14th values. Whichever interval contains these values is called the median group.

2. Find the cumulative percentage of the interval preceding the median group. Label this value **A**.

3. Using this cumulative percentage, calculate how many numbers are needed in order to add up to 50% of the total cumulative percentage. This value will be labeled **B**. Use the following formula to calculate **B**:

   B = 50 - A

4. Figure out the range (how many numbers the interval covers). Call this value **C**. Then, find the percentage for the median interval. Call this value **D**.

5. Calculate how many data values you have to count in the median group to get 50% of the total data set by using the following formula. Call this value **E**.

   E = (B ÷ D) x C

6. Find out what the median value is by adding the value for E to the lower value of the median interval:

   Median = lower value + E

   Since **E = (B ÷ D) x C**, this formula can also be written as:

   Median = lower value + (B ÷ D) x C

If the cumulative frequency for an interval is exactly 50%, then the median value would be the endpoint of this interval.

Let's make this clear with an example!

Example 4 – Grouped variables - frequency distribution (continuous or discrete)

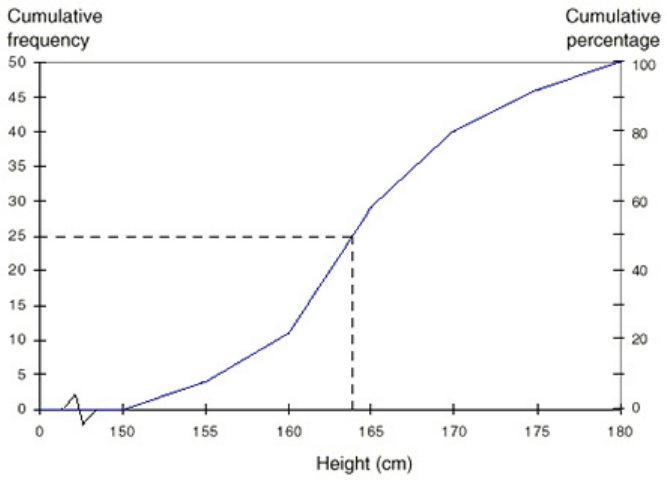
Using the same information from Example 4 in the Mean section, imagine that you surveyed 50 Grade 10 girls to find out how tall each one is in centimetres. After gathering all of your data, you created a frequency distribution table that looked like this:

Table 2. Height of Grade 10 girls

Height (cm)	Frequency (f)	Endpoint (x)	Cumulative frequency	Percentage	Cumulative percentage
150 to < 155	4	155	4	8	8
155 to < 160	7	160	11	14	22
160 to < 165	18	165	29	36	58
165 to < 170	11	170	40	22	80
170 to < 175	6	175	46	12	92
175 to < 180	4	180	50	8	100

Using the grouped data, you created a cumulative frequency graph to accompany your table. The endpoints of the height intervals, the numbers for cumulative frequency and the numbers for cumulative percentage have been plotted on the graph.

Figure 1. Height of Grade 10 girls



By just looking at the graph, you can try to find the median value. The median is the point where the x-axis (Height) intersects with the midpoint (25) of the y-axis (Cumulative frequency). You will see that the median value is approximately 164 cm. Using mathematical calculations, you can find out that the value is actually 163.9 cm. Here's how:

1. According to the information provided in Table 2:

   Median = {(n + 1) ÷ 2}th value
   = (50 + 1) ÷ 2
   = 51 ÷ 2
   = 25.5

   By adding up the frequencies, we find that the median (25.5) lies in the median group of the 160 to < 165 cm interval.

2. The cumulative percentage of the preceding interval (**A**) is 22.

3. The percentage needed in order to get 50% of the total cumulative percentage (**B**) is 28.

   B = 50 - A

= 50 - 22
= 28

4. The range of the median interval (**C**) is 5 and the percentage for the median interval (**D**) is 36.

5. The number of values to count down within the interval in order to get to 50% of the total data set is 3.9.

   E = (B ÷ D) x C
   = (28 ÷ 36) x 5
   = 3.9

6. Since the lower value of the median interval is 160, when you add the value of **E** to that you get a median of 163.9 cm.

   Median = lower value of median interval + (B ÷ D) x C
   = 160 + (28 ÷ 36) x 5
   = 160 + 3.9
   = 163.9 cm

Stem and leaf plots

Ordered stem and leaf plots make it simple to calculate the median, particularly if the cumulative frequencies have already been calculated. Consider the heights of 50 Grade 10 girls using a stem and leaf plot. (See the Organizing data chapter for more information on how to construct these tables.)

Example 5 – Stem and leaf plot

Table 3. Height of Grade 10 girls

Stem* (cm)	Leaf	Cumulative frequency
$15^{(0)}$	0 1 1 4	4
$15^{(5)}$	5 6 7 7 8 8 8	11
$16^{(0)}$	0 1 1 1 1 2 2 2 2 2 2 3 3 3 4 4 4 4	29
$16^{(5)}$	5 5 5 5 6 6 6 7 7 8 9	40
$17^{(0)}$	0 0 1 2 3 3	46
$17^{(5)}$	6 6 7 8	50

***Note:** The stems have been split into smaller intervals. Stem $15^{(0)}$ means that all the data fall within the interval 150 to 154. Stem $15^{(5)}$ means that the data are in the interval of 155 to 159.

There are 50 pieces of data, so the median is the value of the 25.5^{th} observation.

Median = $\{(n + 1) \div 2\}^{th}$ value
= (50 + 1) ÷ 2
= 51 ÷ 2
= 25.5

Therefore, the median lies between the 25^{th} and 26^{th} values. To find out what these values are, count each value in the Leaf column until you have reached the 25^{th} and 26^{th} values. These values lie in the 16(0) interval, meaning the 160 to 164 interval. The numbers in the leaf column represent the numbers in the interval (e.g., 3 represents 163). Thus, the median lies between 163 cm (25^{th} value) and 164 cm (26^{th} value). The median is found by averaging these two values.

Average = (value before median + value after median) ÷ 2
= (25^{th} value + 26^{th} value) ÷ 2
= (163 + 164) ÷ 2
= 327 ÷ 2
= 163.5

Since height is a continuous variable, 163.5 cm is an acceptable median value.

The median obtained from the cumulative frequency graph (164 cm) is not the same value as the median obtained from the calculation in Example 4 (163.9 cm) or from the stem and leaf plot (163.5 cm). This is because you can only find an approximation for the median, unless the graph is drawn precisely with all the information used.

The calculations in Example 4 are only approximations, since grouped data do not tell you how the 36% of the 50 girls found in the median interval are distributed within the interval. As a result, we make the assumption that they are uniformly distributed, and this may lead to a slightly different median. However, the stem and leaf plot is the most accurate means of obtaining a median because it uses all of the actual values.

Comparing the mean and median

It is possible for the mean and median of a distribution to have the same value. This is always the case if distribution is symmetrical as in a normal distribution. If the distribution is roughly symmetrical, then the two values will be close together.

In the example of the heights of the 50 Grade 10 girls, the mean (164.5 cm) is very close to the value of the median (163.5 cm). This is because the distribution is roughly symmetrical (see the stem and leaf plot in the above example).

However, one number can alter the mean without affecting the median.

Example 6 – Comparing the mean and median

Consider the following sets of data that represent the number of points scored by 3 players in 11 lacrosse games.

Eileen: 1, 1, 1, 2, 2, 2, 2, 2, 3, 3, 3
Mean = 22 ÷ 11 = **2**
Median = **2**

Jeremy: 1, 1, 1, 2, 2, 2, 2, 2, 3, 3, 4
Mean = 23 ÷ 11 = **2.1**
Median = **2**

Randy: 1, 1, 1, 2, 2, 2, 2, 2, 3, 3, 14
Mean = 33 ÷ 11 = **3**
Median = **2**

The three sets of data above are identical except for the last observation values (3, 4 and 14).

The median does not alter because it is only dependent on the middle observation's value. The mean does change, however, because it is dependent on the average value of all observations. So, in the above example, as the last value of the last observation increases, so too does the mean.

In the third data set, the value of 14 is very different from any other values. When an observation is very different from all other observations in a data set, it is called an outlier. (For more information on outliers, see the Stem and leaf plots section.) The mean is the measure of central tendency most affected by outliers.

Outliers can sometimes occur as a result of error or deliberate misinformation. In these cases, the outliers should be excluded from the measure of central tendency. Other times, outliers just show how different one value is, and this can be a very useful piece of data.

Example 7 – Comparing the mean and median

When house prices are referred to in newspapers, generally the median price is quoted. Why is this measure used instead of the mean?

There are many moderately priced houses, but there are also some expensive ones and a few very expensive ones. The mean figure could be quite high as it includes the prices of the more expensive houses. But the median gives a more accurate and realistic value of the prices faced by most people.

In summary, the median is the central number and is good to use in skewed (or unbalanced) distributions because it is not affected by outliers.

Example 8 – Comparing the mean and median

Suppose you want to know how much money a family could afford to spend on housing. This would depend on the total family income.

For a family of five (two parents who work and three children with no income) the mean income of each family member is the total income divided by five (e.g., 60,000 ÷ 5 = 12,000). However, the median income would be zero, because more than half of the members of the family make nothing. In some situations, the mean can be much more informative than the median.

Example 9 – Comparing the mean and median

If you want to find out whether a country is wealthy or not, you might consider using the median as your measure of central tendency instead of the mean.

The mean family income could be quite high if income is highly concentrated in a few very wealthy families (despite the fact that most families might earn essentially nothing). Thus, the median family income would be a more meaningful measure—at least half the families would earn the median income or less, and at least half would earn at least as much as the median income or more.

Example 10 – Comparing the mean and median

Suppose you are applying for a job as an accountant at several large firms, and you want to get an idea of how much money you could expect to be making in five years if you join a particular firm. You may want to consider the salaries of accountants in each firm five years after they are hired.

One very high salary could make the mean salary higher; that might not reflect a typical salary within these firms. However, half the accountants make the median salary or less, and half make the median salary or more. So, the measure of central tendency that would give you a better idea of a typical salary would be the median.

Example 11 – Comparing the mean and median

By choosing a measure of central tendency favourable to your point of view, you can mislead people with statistics. In fact, this is commonly done.

Imagine you are the owner of a bakery that makes and sells individual birthday cakes and huge wedding cakes.

It might be in your interest to claim to your customers that the prices have been lowered, and to claim to your shareholders that you have raised the prices. Suppose that last year you sold 100,000 birthday cakes at $10 each, and 1,000 wedding cakes at $1,000 each. This year, you sold 100,000 birthday cakes at $8 each and 1,000 wedding cakes at $1,200 each.

- The median price of the 101,000 cakes sold last year is $10, because more than half of the items sold were birthday cakes. The median price of the 101,000 cakes sold this year is $8.
- The mean price of the 101,000 cakes sold last year is $19.80.

  (100,000 x $10 + 1,000 x $1,000) ÷ 101,000 = $19.80
- The mean price of the 101,000 cakes sold this year is also $19.80.

  (100,000 x $8 + 1,000 x $1,200) ÷ 101,000 = $19.80

The average price per cake sold is the same in both years. Also, the total revenue and the number of the cakes sold was the same. The idea is that you can make data appear to tell conflicting stories by choosing the appropriate measure of central tendency.

It is important to note that you do not have to use only one measure of central tendency. The mean and median can both be used, thus providing more information about the data.

Calculating the mode

In a set of data, the mode is the most frequently observed data value. There may be no mode if no value appears more than any other. There may also be two modes (bimodal), three modes (trimodal), or four or more modes (multimodal). In the case of grouped frequency distributions, the modal class is the class with the largest frequency.

Mode = the most frequently observed data value

As a set of data can have more than one mode, the mode does not necessarily indicate the centre of a data set. The mode will be close to the mean and median if the data have a normal or near-normal distribution. In fact, if the distribution is symmetrical and unimodal, then the mean, the median and the mode may have the same value.

Categorical or discrete variables

For categorical or discrete variables, the mode is simply the most observed value. To work out the mode, observations do not have to be placed in order, although for ease of calculation it is advisable to do so.

Example 1 – Categorical or discrete variables

During a hockey tournament, Anne scored 7, 5, 0, 7, 8, 5, 5, 4, 1, and 5 points in 10 games. The mode of her data set is 5 because this value occurred the most often (four times). This can be interpreted to mean that if one game were selected at random, a good guess would be that Anne would score 5 points.

Example 2 – Categorical or discrete variables

During Marco's 12-game basketball season, he scored 14, 14, 15, 16, 14, 16, 16, 18, 14, 16, 16 and 14 points. This data set is bimodal; there are two modes, 14 and 16, because both of them occur the most often (five times).

Example 3 – Categorical or discrete variables

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

First, put the data set in order:

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

Find and compare the mean, median and mode.

Mode = most frequently observed data value = 0

The mode is 0 because this value occurs most often. If one game were selected at random, the mode tells us that a good guess would be that Jerome would not score a touchdown.

Mean =

Σ

x ÷ n
= 14 ÷ 14
= **1**

However, on average (mean), Jerome will score one touchdown per game even though the mode indicates he did not score a touchdown in a lot of games. In this case, the mode does not provide a useful measure of the player's performance.

Median = (n + 1) ÷ 2^{th} value
= (14 + 1) ÷ 2
= 15 ÷ 2
= 7.5

Average = (value below median + value above median) ÷ 2
= (seventh value + eighth value) ÷ 2
= (1 + 1) ÷ 2
= **1**

Because the number of values in the data set is even, the median does not fit perfectly in the centre of the data set. Instead, the median had to be found using the above equations. According to the results, the median states that Jerome will score one touchdown per game.

Grouped variables (continuous or discrete)

When continuous or discrete variables are grouped in tables, the mode is defined as the class interval where most observations lie. This is called the modal-class interval.

In the example of the height of 50 Grade 10 girls, the *modal-class interval* would be 160 –< 165 cm, as this interval has the most observations in it.

The mode is rarely used as a measure of central tendency for numeric variables. However, for categorical variables, the mode is more useful because the mean and median do not make sense.

Next, you could determine the midrange of the modal class. The *midrange* is simply the midpoint between the highest and lowest values in a class. The mode is not used very often in conjunction with the midrange because it gives only a very poor estimate of the average.

The mode can be used with categorical data, but the mean and median cannot. The mode may or may not exist, and there may be more than one value for the mode.

Summary

Circumstances generally dictate which measure of central tendency—mean, median or mode—is the most appropriate. If you are interested in a total, the mean tends to be the most meaningful measure of central tendency because it is the total divided by the number of data. For example, the mean income of the individuals in a family tells you how much each family member can spend on life's necessities. The median measure is good for finding the central value and the mode is used to describe the most typical case.

Exercises

1. For the following sets of data, find to one decimal place

    i. the mean
    ii. the median, and
    iii. the mode

    a. 0 – 0 – 0 – 0 – 1 – 0 – 0 – 0 – 0 – 0 – 0
    b. 2 – 1 – 2 – 3 – 1 – 3 – 0 – 2 – 4 – 2 – 2
    c. 2.4 – 3.9 – 1.8 – 1.7 – 4.0 – 2.1 – 3.9 – 1.5 – 3.9 – 2.6
    d. 153.8 – 154.7 – 156.9 – 154.3 – 152.3 – 156.1 – 152.3

2. For the following sets of data, find

    i. the mean
    ii. the median
    iii. the mode

    Briefly describe the positions of the mean, median and mode and their relation to one another for each data set.

    a.

    Table 1

x	frequency
-2	3
-1	7
0	8
1	5
2	4

    b.

    Table 2

x	frequency
6.3	2
6.4	1
6.5	6
6.6	5
6.7	13
6.8	4

    c.

    Table 3

x	frequency
1	15
2	5
3	3
4	1
5	2

3. For each of the following stem and leaf plots, find

i. the median, and

ii. the modal-class interval

a.

Table 4

Stem	Leaf
2	2 3 8
3	1 1 4 2
4	2 2 3 5 8 9 9
5	2 4 7 7 8
6	0 3 2
7	4

Stem 4, *Leaf* 2 represents 42

b.

Table 5

Stem	Leaf
$0^{(0)}$	2
$0^{(5)}$	5 6 8
$1^{(0)}$	0
$1^{(5)}$	5 5 6 6 7 8 8 9
$2^{(0)}$	0 0 0 1 1 2 3 3 3 4 4 4
$2^{(6)}$	6 6 7 8 8 9 9
$3^{(0)}$	0 4
$3^{(5)}$	5 6 7 7 8

Stem 3, *Leaf* 5 represents 35

4. Imagine that the annual population increases over a 10 year period are given in the table below:

Table 6. Population increase

Year	Increase from previous
1	53,377
2	52,170
3	67,000
4	90,332
5	72,681
6	65,226
7	76,777
8	83,657
9	77,753
10	82,892

a. Calculate the mean annual population increase over a 10 year period.

b. Calculate the median annual population increase over a 10 year period.

c. Do you think the difference between these two measures is significant? Give reasons for your answer, and explain which result gives a better indication of the data's centre.

d. For what purposes would one use measures such as these?

5. 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.

6. Imagine that the number of unemployed people is given in the table below

Table 7. Unemployment

Age group	No. unemployed
15 to 19	3,688
20 to 24	4,031
25 to 34	5,432
35 to 44	4,360
45 to 54	3,162
55 to 64	1,702

a. Copy the table into your notebook and find the midpoint of each interval. Calculate the average age of an unemployed person using the midpoint.

b. What is the modal-class interval?

c. In what age group does the median lie?

d. Briefly discuss the comparison of these three results.

e. Why do you think the number of unemployed people decreases after the age group 25 to 34?

f. How might social welfare organizations use these figures?

7. A random survey of 100 married men gave the following distribution of hours spent per week doing unpaid household work:

Table 8. Hours spent per week doing unpaid household work

Hours	No. of men
0 to < 5	1
5 to < 10	18
10 to < 15	24
15 to < 20	25
20 to < 25	18
25 to < 30	12
30 to < 35	1
35 to < 40	1

a. Copy the table into your notebook and include columns to find the endpoint (upper value) for each interval. Figure out the cumulative frequency and cumulative percentages and insert them into your table.

b. Draw the ogive (or distribution curve) with the cumulative frequency on the y axis.

c. From the curve, find an approximate median value. What does this value indicate?

d. What is the modal-class interval?

e. Calculate the mean. What does this value indicate?

f. Briefly describe the comparison between the mean, median and mode values.

g. How would you find out whether women spent more hours doing unpaid household work per week than men?

8. The following is a hypothetical table of annual income of people aged 15 years or more:

Table 9. Annual income of people aged 15 years or more

Income ($)	Persons
0 to 2,079	114,195
2,080 to 4,159	44,817
4,160 to 6,239	45,862
6,240 to 8,319	139,611
8,320 to 10,399	114,192
10,400 to 15,599	148,276
15,600 to 20,799	123,638
20,800 to 25,999	121,623
26,000 to 31,199	103,402
31,200 to 36,399	73,463
36,400 to 41,599	59,126
41,600 to 51,999	68,747
52,000 to 77,999	56,710

a. What is the modal-class interval?

b. Copy the table into your notebook and include columns to find the upper endpoint of each interval. Calculate cumulative frequencies and cumulative percentages.

c. Draw the ogive (or distribution curve).

d. From the curve, give an approximate value for the median annual individual income.

e. Calculate the mean annual income. (**Hint**: in the above table, the interval 2,080 to 4,159 actually represents 2,080 to < 4,160, so the midpoint is 3,120.)

f. Briefly compare the mean, median and mode values.

g. Which measure gives the most accurate picture of the data's centre?

h. What types of organization would use information such as this?

Class activities

1. Measure the height of each student in your class to the nearest centimetre. Are there any outliers? Use an appropriate method to find the mean, median and mode. Compare all three measures. Which value gives the best measure of central tendency? Why? Which organizations or companies would find such statistics useful?

2. Find out what your grade or school's student population has been for the last 10 years. Are there any outliers? Use an appropriate method to find the mean, median and mode. Compare all three measures. Which value gives the best measure of central tendency? Why? How would your school or school board use such statistics?

3. Find the final scores of your favourite school sport from your school's records. Collect the scores, both wins and losses, for the last 10 years. (If the data are not available, use data for your favourite sporting team.)
   - What was the mean final score, including both wins and losses, for the past 10 years?
   - What was the median final score, including both wins and losses, for the past 10 years?
   - Are any of the mean final scores similar to the corresponding median final score?
   - Given these values, what can be said about the distributions?
   - What are some of the problems you might come across in trying to use statistics to compare school or other sports teams of the past with those of today?
4. For ordinal data, can you think of occasions where the mode would be of more use than the median or mean? Discuss as a class.

Answers

1.

a.

i. 0.1

ii. 0

iii. 0

b.

i. 2

ii. 2

iii. 2

c.

i. 2.8

ii. 2.5

iii. 3.9

d.

i. 154.3

ii. 154.3

iii. 152.3

2.

a.

i. 0

ii. 0

iii. 0

iv. The mean, median and mode are equal. This distribution is almost symmetrical.

b.

i. 6.6

ii. 6.7

iii. 6.7

iv. Distribution is skewed left, so the mean is less than the median. The mode and median are the same. In skewed distributions, the median is the better measure of central tendency.

c.

i. 1.85

ii. 1

iii. 1

iv. The median and mode are the same. The distribution is skewed right, so the mean is more than the median. In skewed distributions, the median is the better measure of central tendency. In b) and c), the mean has been influenced by a few low and high values.

3.

a.

i. 48

ii. 40 to 49

b.

i. 23

ii. 20 to 24

4.

a. 72,186.5

b. 74,729

c. The measures are quite close together, given the size of each observation. The median probably gives the best indication of the data's centre, as there is a large diversity of observation values. The median would not be affected by the very large or very small values.

d. A government could use these measures when planning for building schools, hospitals and road construction. The government could also use them to help predict revenue intake from taxation.

5.

a.

Table 1. Math test results, marked out of 10 points

Score (x)	Tally	Frequency (f)
0		
1	II	2
2	III	3
3	IIII	4
4	IIII	4
5	IIII	4
6	II	2
7	𝍸 𝍸	10
8	III	3
9	𝍸 I	6
10	II	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.

6.

a. 36.2 to 34 (**Note:** interval sizes are not the same. If they were, the 15 to 24 interval would be the modal-class interval.)

b. 25 to 34

c. All three results lie within the same interval, but distribution is skewed (or slanted) to the right.

d. The younger age groups, 15 to 19 and 20 to 24, are filled with students who are still in school or graduates who have not yet been able to get a job. The age groups after 25 to 34 contain a smaller proportion of unemployed people because these people have joined the work force full time and are no longer attending school.

e. Social welfare organizations might use these figures to plan employment programs catering to younger people.

7.

a.

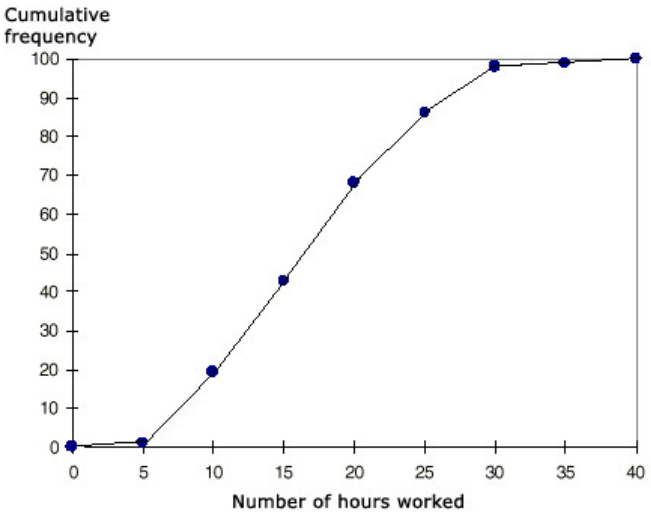
Table 2. Hours spent per week doing unpaid household work

Hours	No. of men (x)	Endpoint	Cumulative frequency	Cumulative percentage
0 to < 5	1	5	1	1
5 to < 10	18	10	19	19
10 to < 15	24	15	43	43
15 to < 20	25	20	68	68
20 to < 25	18	25	86	86
25 to < 30	12	30	98	98

30 to < 35	1	35	99	99
35 to < 40	1	40	100	100

b.

Figure 1. Hours spent per week doing unpaid household work by 100 married men



c. The approximate median value is 17 hours. This indicates that the middle of the distribution is 17 hours.

d. The modal-class interval is 15 to < 20 hours.

e. The mean value is 16.8 hours. This indicates that the average number of hours that a married man spends doing unpaid household work is 16.8 hours.

f. The mean and median are very similar, and all measures lie in the modal-class interval. The distribution is almost symmetrical.

g. A survey could be conducted and analysed in a similar fashion. Then, the results of both surveys could be compared.

8.

a. The modal-class interval is $10,400 to $15,599. (**Note:** interval sizes are not the same.)

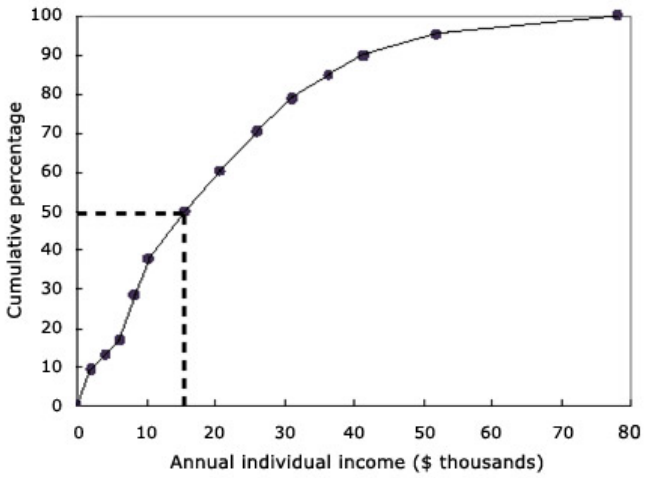
b.

Table 3. Annual income of people aged 15 years and more

Income ($)	Persons	Endpoint	Cumulative frequency	Cumulative percentage
		0	0	0.0
0 to < 2,080	114,195	2,080	114,195	9.4
2,080 to < 4,160	44,817	4,160	159,012	13.1
4,160 to < 6,240	45,862	6,240	204,874	16.9
6,240 to < 8,320	139,611	8,320	344,485	28.4
8,320 to < 10,400	114,192	10,400	458,677	37.8
10,400 to < 15,600	148,276	15,600	606,953	50.0
15,600 to < 20,800	123,638	20,800	730,591	60.2
20,800 to < 26,000	121,623	26,000	852,214	70.2
26,000 to < 31,200	103,402	31,200	955,616	78.7
31,200 to < 36,400	73,463	36,400	1,029,079	84.8
36,400 to < 41,600	59,126	41,600	1,088,205	89.7
41,600 to < 52,000	68,747	52,000	1,156,952	95.3
52,000 to < 78,000	56,710	78,000	1,213,662	100.0

c.

Figure 2. Annual income of people aged 15 years and more



d. The median annual individual income is approximately $15,500.

e. The mean annual income is $19,986.

f. It is difficult to compare the mode with the mean and median because of the difference between the sizes of the intervals. The mean is higher than the median because it is affected by the higher incomes. This means that the distribution is skewed or slanted to the right.

g. The median gives the most accurate picture of the data's centre because it is not influenced by extreme values.

h. Some possible answers include the following:

- social welfare organisations interested in the number of low-income earners;
- businesses interested in the number of high-income earners; and
- governments and other service providers interested in data, broken down by such characteristics as age, sex and geographic area, in order to locate services appropriately.

Measures of spread

Measures of central tendency attempt to identify the most representative value in a set of data. Mean, median and mode give different perspectives of a data set's centre, but a data description is not complete until the spread variability is also known. In fact, the basic numerical description of a data set requires measures of both centre and spread. Measures of spread include range, quartiles, variance and standard deviation.

Range and quartiles

Range

The range is very easy to calculate because it is simply the difference between the largest and the smallest observed values in a data set. Thus, range, including any outliers, is the actual spread of data.

Range = difference between highest and lowest observed values

A great deal of information is ignored when computing the range, since only the largest and smallest data values are considered.

The range value of a data set is greatly influenced by the presence of just one unusually large or small value (outlier).

The range can be expressed as an interval such as 4–10, where 4 is the lowest value and 10 is highest. Often, it is expressed as interval width. For example, the range of 4–10 can also be expressed as a range of 6. The latter convention will be used throughout this chapter.

The disadvantage of using range is that it does not measure the spread of the majority of values in a data set—it only measures the spread between highest and lowest values. As a result, other measures are required in order to give a better picture of the data spread. The range is an informative tool used as a supplement to other measures such as the standard deviation or semi-interquartile range, but it should rarely be used as the only measure of spread.

Quartiles

The median divides the data into two equal sets. For more information on the median, refer to the chapter on Measures of central tendency:

- The lower quartile is the value of the middle of the first set, where 25% of the values are smaller than Q_1 and 75% are larger. This first quartile takes the notation Q_1.
- The upper quartile is the value of the middle of the second set, where 75% of the values are smaller than Q_3 and 25% are larger. This third quartile takes the notation Q_3.

It should be noted that the median takes the notation Q_2, the second quartile.

Example 1 – Upper and lower quartiles

- Data: 6, 47, 49, 15, 43, 41, 7, 39, 43, 41, 36
- Ordered data: 6, 7, 15, 36, 39, 41, 41, 43, 43, 47, 49
- Median: 41
- Upper quartile: 43
- Lower quartile: 15

Interquartile range

The interquartile range is another range used as a measure of the spread. The difference between upper and lower quartiles (Q_3–Q_1), which is called the interquartile range, also indicates the dispersion of a data set. The interquartile range spans 50% of a data set, and eliminates the influence of outliers because, in effect, the highest and lowest quarters are removed.

Interquartile range = difference between upper quartile (Q_3) and lower quartile (Q_1)

Example 2 – Range and quartiles

A year ago, Angela began working at a computer store. Her supervisor asked her to keep a record of the number of sales she made each month.

The following data set is a list of her sales for the last 12 months:

34, 47, 1, 15, 57, 24, 20, 11, 19, 50, 28, 37.

Use Angela's sales records to find:

a. the median

b. the range

c. the upper and lower quartiles

d. the interquartile range

Answers

a. The values in ascending order are:
1, 11, 15, 19, 20, 24, 28, 34, 37, 47, 50, 57.

Median = (12th + first) ÷ 2
= 6.5th value
= (sixth + seventh observations) ÷ 2
= (24 + 28) ÷ 2
= **26**

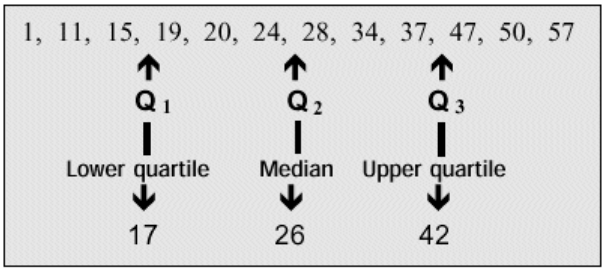
b. Range = difference between the highest and lowest values
= 57 – 1
= **56**

c. Lower quartile = value of middle of first half of data Q_1
= the median of 1, 11, 15, 19, 20, 24
= (third + fourth observations) ÷ 2
= (15 + 19) ÷ 2
= **17**

d. Upper quartile = value of middle of second half of data Q_3
= the median of 28, 34, 37, 47, 50, 57
= (third + fourth observations) ÷ 2
= (37 + 47) ÷ 2
= **42**

e. Interquartile range = Q_3–Q_1
= 42 – 17
= **25**

These results can be summarized as follows:



Note: This example has an even number of observations. The median, Q_2, lies between the centre of two observations (24 and 28), so the calculation of Q_1 includes the observation 24 as it is below the value of Q_2. Similarly, 28 is also included in the calculation of Q_3 as it is above the value of Q_2.

Consider an odd number of observations such as 1, 2, 3, 4, 5, 6, 7. Here the value of Q_2 is 4. As the location of the median is right on the fourth observation, this value is not included in calculating Q_1 and Q_3 , as we are interested only in the data above and below Q_2. In the above example, Q_1 = 2 and Q_3 = 6.

Semi-quartile range

The semi-quartile range is another measure of spread. It is calculated as one half the difference between the 75th percentile (often called Q_3) and the 25th percentile (Q_1). The formula for semi-quartile range is:

(Q_3–Q_1) ÷ 2.

Since half the values in a distribution lie between Q_3 and Q_1, the semi-quartile range is one-half the distance needed to cover half the values. In a symmetric distribution, an interval stretching from one semi-quartile range below the median to one semi-quartile above the median will contain one-half of the values. However, this will not be true for a skewed distribution.

The semi-quartile range is hardly affected by higher values, so it is a good measure of spread to use for skewed distributions, but it is rarely used for data sets that have normal distributions. In the case of a data set with a normal distribution, the standard deviation is used instead.

Variance and standard deviation

Unlike range and quartiles, the variance combines all the values in a data set to produce a measure of spread. The variance (symbolized by $\mathbf{S^2}$) and standard deviation (the square root of the variance, symbolized by **S**) are the most commonly used measures of spread.

We know that variance is a measure of how spread out a data set is. It is calculated as the average squared deviation of each number from the mean of a data set. For example, for the numbers 1, 2, and 3 the mean is 2 and the variance is 0.667.

$[(1 - 2)^2 + (2 - 2)^2 + (3 - 2)^2] \div 3 = 0.667$

[squaring deviation from the mean] ÷ number of observations = variance

Variance (S^2) = average squared deviation of values from mean

Calculating variance involves squaring deviations, so it does not have the same unit of measurement as the original observations. For example, lengths measured in metres (m) have a variance measured in metres squared (m^2).

Taking the square root of the variance gives us the units used in the original scale and this is the standard deviation.

Standard deviation (S) = square root of the variance

Standard deviation is the measure of spread most commonly used in statistical practice when the mean is used to calculate central tendency. Thus, it measures spread around the mean. Because of its close links with the mean, standard deviation can be greatly affected if the mean gives a poor measure of central tendency.

Standard deviation is also influenced by outliers one value could contribute largely to the results of the standard deviation. In that sense, the standard deviation is a good indicator of the presence of outliers. This makes standard deviation a very useful measure of spread for symmetrical distributions with no outliers.

Standard deviation is also 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.

Generally, the more widely spread the values are, the larger the standard deviation is. For example, imagine that we have to separate two different sets of exam results from a class of 30 students the first exam has marks ranging from 31% to 98%, the other ranges from 82% to 93%. Given these ranges, the standard deviation would be larger for the results of the first exam.

Standard deviation might be difficult to interpret in terms of how big it has to be in order to consider the data widely spread. The size of the mean value of the data set depends on the size of the standard deviation. When you are measuring something that is in the millions, having measures that are "close" to the mean value does not have the same meaning as when you are measuring the weight of two individuals. 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 useful to assess the size of the standard deviation relative to the mean of the data set.

Although standard deviation is less susceptible to extreme values than the range, standard deviation is still more sensitive than the semi-quartile range. If the possibility of high values (outliers) presents itself, then the standard deviation should be supplemented by the semi-quartile range.

Properties of standard deviation

When using standard deviation keep in mind the following properties.

- Standard deviation is only used to measure spread or dispersion around the mean of a data set.
- Standard deviation is never negative.
- Standard deviation is sensitive to outliers. A single outlier can raise the standard deviation and in turn, distort the picture of spread.
- For data with approximately the same mean, the greater the spread, the greater the standard deviation.
- If all values of a data set are the same, the standard deviation is zero (because each value is equal to the mean).

When analysing normally distributed data, standard deviation can be used in conjunction with the mean in order to calculate data intervals.

If $\bar{\mathbf{x}}$ = mean, **S** = standard deviation and **x** = a value in the data set, then

- about 68% of the data lie in the interval: $\bar{\mathbf{x}} - \mathbf{S} < \mathbf{x} < \bar{\mathbf{x}} + \mathbf{S}$.
- about 95% of the data lie in the interval: $\bar{\mathbf{x}} - \mathbf{2S} < \mathbf{x} < \bar{\mathbf{x}} + \mathbf{2S}$.

- about 99% of the data lie in the interval: $\bar{x} - 3S < x < \bar{x} + 3S$.

Discrete variables

The variance for a discrete variable made up of **n** observations is defined as:

$$S^2 = \frac{\sum (x - \bar{x})^2}{n}$$

The standard deviation for a discrete variable made up of **n** observations is the positive square root of the variance and is defined as:

$$S = \sqrt{\frac{\sum (x - \bar{x})^2}{n}}$$

Use this step-by-step approach to find the standard deviation for a discrete variable.

1. Calculate the mean.
2. Subtract the mean from each observation.
3. Square each of the resulting observations.
4. Add these squared results together.
5. Divide this total by the number of observations (variance, S^2).
6. Use the positive square root (standard deviation, **S**).

Example 1 – Standard deviation

A hen lays eight eggs. Each egg was weighed and recorded as follows:

60 g, 56 g, 61 g, 68 g, 51 g, 53 g, 69 g, 54 g.

a. First, calculate the mean:

$$\bar{X} = \frac{\sum x}{n}$$
$$= \frac{472}{8}$$
$$= 59$$

b. Now, find the standard deviation.

Table 1. Weight of eggs, in grams

Weight (x)	$(x - \bar{x})$	$(x - \bar{x})^2$
60	1	1
56	-3	9
61	2	4
68	9	81
51	-8	64
53	-6	36
69	10	100
54	-5	25
472		**320**

Using the information from the above table, we can see that

$$\sum (x - \bar{x})^2 = 320$$

In order to calculate the standard deviation, we must use the following formula:

$$S = \sqrt{\frac{\sum (x - \bar{x})^2}{n}}$$
$$= \sqrt{\frac{320}{8}}$$
$$= 6.32 \text{ grams}$$

Frequency table (discrete variables)

The formulas for variance and standard deviation change slightly if observations are grouped into a frequency table. Squared deviations are multiplied by each frequency's value, and then the total of these results is calculated.

In a frequency table, the variance for a discrete variable is defined as

$$S^2 = \frac{\sum (x - \bar{x})^2 f}{n} \quad \text{where: } n = \sum f$$

The standard deviation for a discrete variable is defined as

$$S = \sqrt{\frac{\sum (x - \bar{x})^2 f}{n}}$$

Example 2 – Standard deviation calculated using a frequency table

Thirty farmers were asked how many farm workers they hire during a typical harvest season. Their responses were:

4, 5, 6, 5, 3, 2, 8, 0, 4, 6, 7, 8, 4, 5, 7, 9, 8, 6, 7, 5, 5, 4, 2, 1, 9, 3, 3, 4, 6, 4

Table 2. Thirty farmers were asked how many farm workers they hire during a typical harvest season. Their responses were:

Workers (x)	Tally	Frequency (f)	(xf)	$(x - \bar{x})$	$(x - \bar{x})^2$	$(x - \bar{x})^2 f$
0	I	1	0	-5	25	25
1	I	1	1	-4	16	16
2	II	2	4	-3	9	18
3	III	3	9	-2	4	12
4	𝍸 I	6	24	-1	1	6
5	𝍸	5	25	0	0	0
6	IIII	4	24	1	1	4
7	III	3	21	2	4	12
8	III	3	24	3	9	27
9	II	2	18	4	16	32
		30	**150**			**152**

To calculate the mean:

$$\bar{x} = \frac{\sum x f}{\sum f}$$
$$= \frac{150}{30}$$
$$= 5$$

To calculate the standard deviation:

$$S = \sqrt{\frac{\sum (x - \bar{x})^2 f}{n}}$$
$$= \sqrt{\frac{152}{30}}$$
$$= 2.25$$

Example 3 – Standard deviation using grouped variables (continuous or discrete)

220 students were asked the number of hours per week they spent watching television. With this information, calculate the mean and standard deviation of hours spent watching television by the 220 students.

Table 3. Number of hours per week spent watching television

Hours	Number of students
10 to 14	2
15 to 19	12
20 to 24	23
25 to 29	60
30 to 34	77
35 to 39	38
40 to 44	8

a. First, using the number of students as the frequency, find the midpoint of time intervals.

b. Now calculate the mean using the midpoint (**x**) and the frequency (**f**).

Note: In this example, you are using a continuous variable that has been rounded to the nearest integer. The group of **10 to 14** is actually 9.5 to 14.499 (as the 9.5 would be rounded up to 10 and the 14.499 would be rounded down to 14). The interval has a length of 5 but the midpoint is 12 (9.5 + 2.5 = 12).

$$\bar{x} = \frac{\sum x f}{\sum f}$$
$$= \frac{6,560}{220}$$
$$= 29.82$$

6,560 = (2 X 12 + 12 X 17 + 23 X 22 + 60 X 27 + 77 X 32 + 38 X 37 + 8 X 42)

Then, calculate the numbers for the **xf**, $\mathbf{(x - \bar{x})}$, $\mathbf{(x - \bar{x})^2}$ and $\mathbf{(x - \bar{x})^2 f}$ formulas.

Add them to the frequency table below.

Table 4. Number of hours spent watching television

Hours	Midpoint (x)	Frequency (f)	xf	$(x - \bar{x})$	$(x - \bar{x})^2$	$(x - \bar{x})^2 f$
10 to 14	12	2	24	-17.82	317.6	635.2
15 to 19	17	12	204	-12.82	164.4	1,972.8
20 to 24	22	23	506	-7.82	61.2	1,407.6
25 to 29	27	60	1,620	-2.82	8.0	480.0
30 to 34	32	77	2,464	2.18	4.8	369.6
35 to 39	37	38	1,406	7.18	51.6	1,960.8
40 to 44	42	8	336	12.18	148.4	1,187.2
		220	**6,560**			**8,013.2**

Example 4 – Standard deviation

Use the information found in the table above to find the standard deviation.

$$S = \sqrt{\frac{\sum (x - \bar{x})^2 f}{n}}$$
$$= \sqrt{\frac{8,013.2}{220}}$$
$$= \sqrt{36.42}$$
$$= 6.03$$

Note: During calculations, when a variable is grouped by class intervals, the midpoint of the interval is used in place of every other value in the interval. Thus, the spread of observations within each interval is ignored. This makes the standard deviation *always* less than the true value. It should, therefore, be regarded as an approximation.

Example 5 – Standard deviation

Assuming the frequency distribution is approximately normal, calculate the interval within which 95% of the previous example's observations would be expected to occur.

$\bar{x}$ = 29.82, **s** = 6.03

Calculate the interval using the following formula: $\bar{x}$ - **2s** < **x** < $\bar{x}$ + **2s**

29.82 - (2 X 6.03) < **x** < 29.82 + (2 X 6.03)

29.82 - 12.06 < **x** < 29.82 + 12.06

17.76 < x < 41.88

This means that there is about a 95% certainty that a student will spend between 18 hours and 42 hours per week watching television.

Five-number summaries

A five-number summary is especially useful in descriptive analyses or during the preliminary investigation of a large data set. A summary consists of five values: the most extreme values in the data set (the maximum and minimum values), the lower and upper quartiles, and the median. These values are presented together and ordered from lowest to highest: minimum value, lower quartile (Q_1), median value (Q_2), upper quartile (Q_3), maximum value.

These values have been selected to give a summary of a data set because each value describes a specific part of a data set: the median identifies the centre of a data set; the upper and lower quartiles span the middle half of a data set; and the highest and lowest observations provide additional information about the actual dispersion of the data. This makes the five-number summary a useful measure of spread.

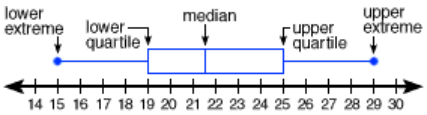
A five-number summary can be represented in a diagram known as a box and whisker plot. In cases where we have more than one data set to analyse, a five-number summary with a corresponding box and whisker plot is constructed for each.

To illustrate the five-number summary of Example 2 in the Range and quartiles section would be 1, 17, 26, 42, 57.

Constructing box and whisker plots

A box and whisker plot (sometimes called a boxplot) is a graph that presents information from a five-number summary. It does not show a distribution in as much detail as a stem and leaf plot or histogram does, but is especially useful for indicating whether a distribution is skewed and whether there are potential unusual observations (outliers) in the data set. Box and whisker plots are also very useful when large numbers of observations are involved and when two or more data sets are being compared. (See the section on five-number summaries for more information.)

Box and whisker plots are ideal for comparing distributions because the centre, spread and overall range are immediately apparent.

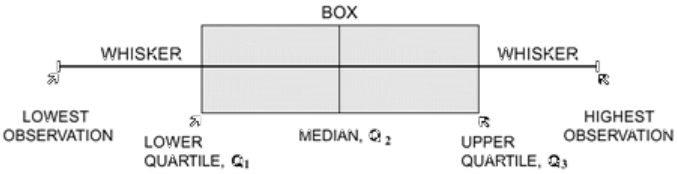


A box and whisker plot is a way of summarizing a set of data measured on an interval scale. It is often used in explanatory data analysis. This type of graph is used to show the shape of the distribution, its central value, and its variability.

In a box and whisker plot:

- the ends of the box are the upper and lower quartiles, so the box spans the interquartile range
- the median is marked by a vertical line inside the box
- the whiskers are the two lines outside the box that extend to the highest and lowest observations.

Figure 1. Box and whisker plot



Example 1 – Box and whisker plots

Like Angela, Carl works at a computer store. He also recorded the number of sales he made each month. In the past 12 months, he sold the following numbers of computers:

51, 17, 25, 39, 7, 49, 62, 41, 20, 6, 43, 13.

1. Give a five-number summary of Carl's and Angela's sales.
2. Make two box and whisker plots, one for Angela's sales and one for Carl's.
3. Briefly describe the comparisons between their sales.

Answers

1. First, put the data in ascending order. Then find the median.

   6, 7, 13, 17, 20, 25, 39, 41, 43, 49, 51, 62.
   Median = (12th + 1st) ÷ 2 = 6.5th value
   = (sixth + seventh observations) ÷ 2
   = (25 + 39) ÷ 2
   = **32**

   There are six numbers below the median, namely: 6, 7, 13, 17, 20, 25.
   Q1 = the median of these six items
   = (6 + 1) ÷ 2= 3.5^{th} value
   = (third + fourth observations) ÷ 2
   = (13 + 17) ÷ 2
   = **15**

   Here are six numbers above the median, namely: 39, 41, 43, 49, 51, 62.
   Q_3 = the median of these six items
   = (6 + 1) ÷ 2= 3.5^{th} value
   = (third + fourth observations) ÷ 2

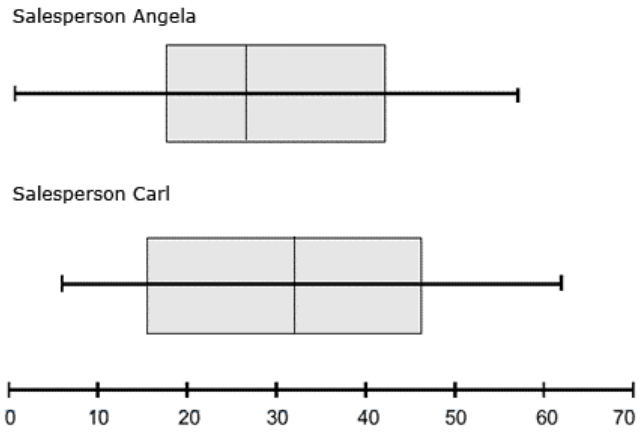
= **46**

The five-number summary for Carl's sales is 6, 15, 32, 46, 62.

Using the same calculations, we can determine that the five-number summary for Angela is 1, 17, 26, 42, 57.

2. Please note that box and whisker plots can be drawn either vertically or horizontally.

Figure 2. Carl's and Angela's box and whisker plots



3. Carl's highest and lowest sales are both higher than Angela's corresponding sales, and Carl's median sales figure is higher than Angela's. Also, Carl's interquartile range is larger than Angela's.

   These results suggest that Carl consistently sells more computers than Angela does.

Summary

There are several ways to describe the centre and spread of a distribution. One way to present this information is with a five-number summary. It uses the median as its centre value and gives a brief picture of the other important distribution values. Another measure of spread uses the mean and standard deviation to decipher the spread of data. This technique, however, is best used with symmetrical distributions with no outliers.

Despite this restriction, the mean and standard deviation measures are used more commonly than the five-number summary. The reason for this is that many natural phenomena can be approximately described by a normal distribution. And for normal distributions, the mean and standard deviation are the best measures of centre and spread respectively.

Standard deviation takes every value into account, has extremely useful properties when used with a normal distribution, and is mathematically manageable. But the standard deviation is not a good measure of spread in highly skewed distributions and, in these instances, should be supplemented by other measures such as the semi-quartile range.

The semi-quartile range is rarely used as a measure of spread, partly because it is not as manageable as others. Still, it is a useful statistic because it is less influenced by extreme values than the standard deviation, is less subject to sampling fluctuations in highly skewed distributions and is limited to only two values Q_1 and Q_3. However, it cannot stand alone as a measure of spread.

Exercises

1. For the following sets of data, find the range.
    a. 6, 8, 11, 15, 24, 38
    b. 11, -6, -2, 16, 9, -8, 17, 19
    c. 6.4, 3.8, 5.9, 4.7, 5.3, 7.1, 3.2

2. Imagine that the number of marriages registered over a 10 year period were as follows:

Table 1. Number of registered marriages

Year	Number of marriages
1	40,650
2	40,812
3	41,300
4	41,450
5	39,594
6	40,734
7	39,993
8	38,814
9	37,828
10	35,716

Using the information from the above table, find
    a. the range
    b. the median
    c. the upper and lower quartiles
    d. the interquartile range
    e. the five-number summary

3. Listed below are the maximum daily temperatures (in degrees Celsius) in Iqaluit from June 2 to June 16:

2.8, 7.3, 9.6, 8.9, 11.4, 6.7, 5.8, 5.5, 6.7, 6.2, 9.0, 8.2, 7.6, 8.5, 6.7
    a. Find the range.
    b. Calculate the interquartile range.
    c. Calculate the five-number summary.

4. The following table outlines hypothetical numbers of industrial disputes over a ten year period;

Table 2. Number of industrial disputes

Year	Number of industrial disputes
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.

5. 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 frequency distribution table.

b. Draw a vertical bar graph.

c. Calculate the mean, median and mode.

d. Calculate the variance and standard deviation.

e. Calculate the interval within which 95% of observations would be expected to occur.

f. Comment on the spread of the data.

Answers

Note: There are several ways to calculate quartiles. If you are using SAS or EXCEL software, the results for these exercises may vary from the answers provided here.

1.

a. 32

b. 27

c. 3.9

2.

a. 5,734

b. 40,321.5

c. Q_1 = 38,814 Q_3 = 40,812

d. 1,998

e. 35,716, 38,814, 40,321.5, 40,812, 41,450

3.

a. 8.6

b. 2.7

c. 2.8, 6.2, 7.3, 8.9, 11.4

4.

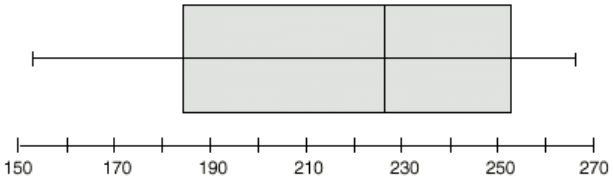
a. 113

b. 78

c. 153, 182, 226.5, 260, 266

d.

Figure 1. Number of industrial disputes



5.

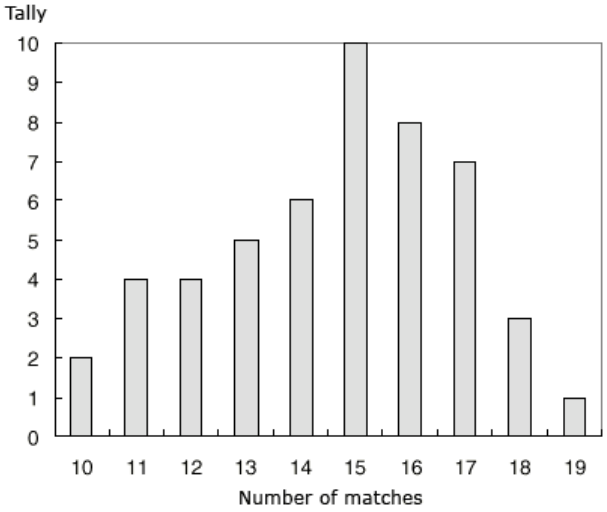
a.

Table 1. Number of basketball games attended by season ticket-holders

Number of matches (x)	Tally	Frequency (f)
10	II	2
11	IIII	4
12	IIII	4
13	𝍸	5
14	𝍸 I	6
15	𝍸 𝍸	10
16	𝍸 III	8
17	𝍸 II	7
18	III	3
19	I	1
		50

b.

Figure 2. Number of basketball games attended by season ticket-holders



c. mean = 14.62, median = 15, mode = 15

d. $\mathbf{S}^2 = 4.96$, $\mathbf{S} = 2.23$

e. $10.16 < \mathbf{x} < 19.08$

f. The standard deviation is quite low, which indicates that the data is not widely spread about the mean. The mean and median are very close together, which indicates that the data are roughly symmetrical.

Information: Use in society

You have probably heard of the term *the information age*. It describes how modern society depends on information as a resource. People rely on information to make decisions and recommendations in many fields, including politics, economics, environment and entertainment.

This chapter will show how people and organizations use statistical information. Without reliable information, people can make poor decisions that sometimes result in serious consequences.

Using information

There are a few basic principles that need to be followed when using statistical data. Keep in mind the following guidelines:

- Find out the facts. Identify which organization or sources can help you find the information you require.
- Decide which format suits your needs best. Information can be released in many forms, including print and online publications, CD-ROMs and more and more from Internet sites.
- Find the information. If Statistics Canada is your source, you may obtain data on the Statistics Canada website, and from most community libraries. If you use other sources, ensure that the information is up-to-date and reliable. Reference librarians across the country will be able to assist you.
- Understand the terms, concepts and definitions of the data you are using, including the surveyed population and the methodology used. This will help you to compare similar information from different sources. Common statistical terms include: current and constant dollars, indexes, random rounding, and seasonal adjustment.

Here are just a few examples of the practical uses of statistical information:

- **Tracking market share:** A Halifax trust and mortgage loan company wants to evaluate its market share of retirement savings plans, mortgages and other key financial products. Using balance sheet data, researchers for the company can compare their growth and market share with the industry total and with other financial groups who sell the same products. By doing this, they define the products that offer the most potential for new sales. They can then integrate these vital facts into a new marketing strategy.
- **Projecting construction costs:** A hydro utility in Yellowknife is planning to build a new generating plant, but it will not be complete for at least 10 years. To establish its viability, planners want to look at future construction costs. The generating plant will be complex, with industrial, commercial and institutional sections. Through consultations with Statistics Canada, planners can determine which indexes they can use and (assuming stability in the economy) project price changes for the next 10 years.
- **Tracking sales prices:** A pharmaceutical firm in Winnipeg wants to compare price changes for its products with those of the industry as a whole. Using a series of pharmaceutical indexes, analysts can compare their price trends over time with those of the industry. In this way, they can measure their own competitiveness.
- **Forecasting a market:** Machine toolmakers in Sudbury need to know what the automotive industry is planning to spend on machinery and equipment. Using capital expenditures data, they can obtain the information they need to plan investments in new plants and equipment.
- **Planning diversification:** A Montréal clothing wholesaler has been doing exceptionally well and wants to expand into another service industry in the province. The wholesaler first looks at retail data to find out where consumer spending is highest, not only throughout the province but for Montréal as well. The index also shows retail trends over time. Using this information, the wholesaler can decide which area offers the most business potential.
- **A major acquisition:** Executives of a British Columbia resource corporation are reviewing a number of corporations for a possible acquisition. They need to collect information about their legal situation, corporate structures and subsidiaries, as well as the names of all directors and officers of each corporation. They also want to study the structure of several competitors to see how integrated they are within the industry. Working closely with the officials of the *Corporation Returns Acts*, they access the Intercorporate Ownership Database. The database provides a list of all corporate holdings and principal industries, as well as the names and addresses of all directors and officers.

There are many more uses of statistical data. To find out what these uses are, proceed to the case studies.

Case study: Ozone layer depletion and the Montréal Protocol

The ozone layer is an important part of the global atmosphere and climate system. It limits the amount of ultraviolet (UV) radiation from the sun to levels necessary for life on Earth. A depleted ozone layer may likely cause serious consequences including higher rates of sunburn, skin cancer, eye damage and other diseases, as well as reducing plant growth.

Manufactured chemical compounds are the main cause of ozone layer depletion. These are compounds such as chlorofluorocarbons (CFCs) and halons, among others. In the past, these compounds were commonly used in refrigerators, air conditionners and fire-retardant chemicals. In general, when atmospheric ozone falls 1%, it is equivalent to an increase of 1%–2% in UV radiation at ground level.

State of the ozone layer

Since 1979, stratospheric ozone has decreased over the entire globe—between 4% and 6% per decade in mid-latitudes and between 10% and 12% per decade in higher southern latitudes. The levels dropped to record lows following the June 1991 volcanic eruption of Mount Pinatubo in the Philippines. However, the effects of this natural disaster have diminished, and levels have returned to values closer to the long-term downward trend.

Potential effects of ozone depletion

Stratospheric ozone depletion leads to increases in UV radiation reaching the earth's surface. High levels of UV radiation are known to slow plant growth. They may also lead to skin cancers, cataracts and immunosuppressive diseases in humans and other animals. At mid-latitudes—for example, where Toronto is located—under clear skies, a 1% decrease in the thickness of the stratospheric ozone layer results in about a 1.1% to 1.4% increase in UV-B at ground level. This varies according to the season. In Canada, about 200 species of crops and trees are, to some degree, sensitive to increased levels of UV-B.

Responses

As one of the original parties to the 1987 Montréal Protocol, Canada has taken a leadership role both in understanding the science behind ozone depletion and in acting to eliminate its causes. The production of ozone-depleting substances (ODS) in Canada has dropped from a high of 27.8 kilotonnes in 1987 to 1.0 kilotonne in 1996. At the global level, 1995 production of CFCs was 77% lower than its peak in 1988. Canada accounted for less than 1% of global production.

Despite this progress, there are still concerns. First, scientists cannot be certain that, even with current elimination targets of ODSs, the ozone layer will return to its previous thickness. The concentration of known ODSs in the stratosphere is declining, although there may be other substances that are contributing to ozone depletion. Second, developing countries now represent the biggest threat to the recovery of the ozone layer, as their production and use of CFCs has grown in recent years. Third, there is a 'loss of momentum' in developed countries, including Canada, because of the perception that the problem has been resolved. The 1997 report of the Auditor General of Canada indicates that current stocks of ODSs, those in existing equipment and quantities stored for future use, are at risk of being released to the atmosphere. This may happen unless more stringent inspections and safeguards of those stocks improve the recovery of the ozone layer by more than 10%.

The problem of ozone layer depletion became prominent in the 1980s. Scientific measurements began to show significant global decreases in ozone. Some of the general results follow.

- For mid-latitudes, Europe and North America, annual ozone losses of 2% to 4% over the 1980s were reported.
- For Australia, ozone losses during the 1980s ranged from 0.5% to 5%.
- For Antarctica, the ozone hole has become a regular feature of the southern hemisphere with total ozone losses of 60% to 70% reported each spring since 1985.

The seriousness of the problem has led to global agreement to reduce and control the production of ODSs. In 1987, 149 countries gathered in Montréal and signed an agreement to reduce the use of ODSs. The decisions taken were the following:

- Freeze consumption of CFC-11 at 1986 levels by 1989.
- Reduce consumption of CFC-12 by 20% by 1 July 1993.
- Make an effort to meet Montréal targets as seen below.

Table 1: Ozone depleting substances

	Ozone depletion potential[1]	Canada's phase-out date	Lifetime in atmosphere
Halons	3.0 to 10.0	Jan. 1, 1994	up to 65 years
Carbon tetrachloride	1.1	Jan. 1, 1995	up to 42 years
CFCs	0.6 to 1.0	Jan. 1, 1996	from 50 to 1,700 years
Methyl chloroform	0.1	Jan. 1, 1996	6 years
Methyl bromide	0.6	Jan. 1, 2005[2]	up to 2 years
HCFCs[3]	0.001 to 0.52	Jan. 1, 2020	up to 19 years

Notes:

1. Each substance that affects the ozone layer is measured using a standardized reference known as the ozone depletion potential (ODP). For further information, see Environment Canada, 1997, *Stratospheric Ozone Depletion, National Environmental Indicator Series, SOE bulletin No. 97-2,* Ottawa.

2. The regulations of the *Environmental Protection Act* require that ODSs must be phased out by January 1, 2010. Methyl bromide will be phased out by developed countries

party to the Montréal Protocol by 2005.

3. Most hydrochlorofluorocarbons have been developed for use as transitional chemicals to replace the more damaging ODSs, mainly CFCs.

Source: Auditor General of Canada. Accessed June 14, 1999. "Ozone Layer Protection: The Unfinished Journey." *Report of the Auditor General of Canada to the House of Commons 1997.*

Case study: Border crossings by car from the United States to Canada: What drove people away?

In this activity, you will compare the number of automobiles entering Canada from the United States via the Fort Erie International Bridge for the years 2000 and 2001. During the first part of 2000, there were fewer crossings than in the first part of 2001. However, the border crossings between September through December 2001 were less than the same time period in the previous year.

Your main goal in this case study is to determine why there was such a large difference in the number of crossings, particularly in the months of September through December. How were events in the early part of 2000 different from those in the early part of the year 2001? Was the weather bad? Was the price of gasoline high? Before continuing, brainstorm some other reasons that might explain the difference.

Although difficult to do, it is critical that anyone who studies data should first question the methods used to make the calculations. If a new procedure to count cars has replaced the one used in previous years, then this could be the source of the varying numbers. Even if you assume that the same counting methods have been in place for both years, you should still question the counting methods used in the future.

Exercise 1

For this exercise, follow the directions below to access the data you will be using for this case study.

1. Log onto the E-STAT website.
2. Under the **People** section, click on the link for **Travel and Tourism**.
3. When the new page opens, select **International Travel** from the folder list.
4. Click on **Table 427-0002 (Number of vehicles travelling between Canada and the United States, monthly).**
5. Choose the following options:

- Geography: select **Fort Erie, Ontario**.
- Trip characteristics: select **Total United States vehicles entering**.
- Length of stay: select **Length of stay, total**.
- Mode of transportation: select **Automobiles**.
- Time: select from **Jan. 2000** to **Dec. 2001**.

- Select Option 2 **Retrieve as individual Time series**.
- Under **OUTPUT SCREEN formats**, choose the option **Plain text table, time as rows**. Leave everything else as it is. Click on **Retrieve now**.
- Record the results in Table 1.

Table 1
Total number of automobiles entering Fort Erie, Ontario per month from the United States (all lengths of stay) 2000 and 2001

	2000	2001
January	126,847	142,810
February		
March		
April		
May		
June		
July		
August		
September		
October		
November		
December		

Exercise 2

Using the results from Table 1, calculate the actual change by subtracting the 2000 value from the 2001 value. Then calculate the percent change in the total number of automobiles that entered Fort Erie for each month from 2000 to 2001 by expressing the amount of change as a percent of the 2000 value. Record the results in Table 2.

Table 2
Monthly change and percent change in the total number of automobiles entering Fort Erie, Ontario from the United States (all lengths of stay)

	2000	2001
January	15,963	13.6
February		
March		
April		
May		
June		
July		
August		
September		
October		
November		
December		

Exercise 3

Study the results in Table 2 and verify whether there was a significant difference between the last four months of 2000 and the same period in 2001. In both years, which of these four months shows the biggest drop in border crossings by car?

Exercise 4

The price of gasoline may be one possible explanation for the differences noted and verified in Exercise 3. To find out whether gas prices were higher in this time interval, visit Statistics Canada's E-STAT website at http://estat.statcan.ca/content/english/over.shtml. Navigate your way through to data for fuel prices. Could gas prices have influenced the lower number of cars entering Canada from the United States in the last four months of 2000?

Exercise 5

Using data from the Statistics Canada website, investigate whether the exchange rate for American to Canadian currency contributed to the lower number of cars entering Canada in the last four months of 2000. A suggested path to accessing this data is given below. You are welcome to find your own route.

At the Statistics Canada Home page, click **Summary Tables** on the left bar.

Find the table **Exchange rates, interest rates, money supply and stock prices** (http://www40.statcan.ca/l01/cst01/econ07.htm).

Click on the link for the Bank of Canada at the bottom of the table.

Once the page is opened, select **English**, then select **More exchange rates.**

Follow the link to the Monthly Average Exchange Rates page.

Many exchange rates are available. To better understand the low number of border crossings, you may want to study the noon exchange rates for the U. S. dollar from 2000 to 2001.

Exercise 6

Another possible explanation for the lower number of border crossings might be that the weather in the later part of 2001 was terrible compared with the weather at the same time in 2000. Visit the Environment Canada site. Determine whether the weather played a role in the lower numbers or could there be other unusual factors that contributed to the change (i.e., a historic event).

Exercise 7

Make a list of other reasons why the numbers might be lower. Using data collected from the Statistics Canada website and possibly other sites, determine whether these factors could have influenced the low numbers.

Nota :

You can also use E-STAT to confirm your calculations of the percent change from 2000 to 2001. Under the **SCREEN OUTPUT formats**, and after the text **'The output will contain',** click the down arrow and select **'percent changes, year-to-year'**.

Case study: Do you feel uneasy about e-banking?

Despite the increase in the number of Canadians surfing the Internet over the last few years, there are still many people who have strong reservations about *e-banking* (electronic banking) and *e-commerce* (purchasing goods and services online). The reason for this is often because people feel insecure about entering their personal information into such a publicly accessible tool as the World Wide Web (WWW). Strangely, however, many of these same people have no hesitations about sending e-mail messages containing personal or confidential information.

In this case study, we will examine data on differing types of Internet activity and predict how these numbers will change in the near future. We will consider some of the security measures used to prevent unlawful access to personal information, as well as take a look at the tools used to make the transfer of confidential information both safe and secure. A closer examination and understanding of these security measures might encourage more people to use the Internet for the purpose of financial transactions.

Exercise 1 – Average household percentage of Internet users

1. Based on your own personal experience and that of your family and friends, estimate the percentage of Canadian households who use the Internet for the purpose of e-mail messaging, e-banking or e-commerce. Keep a record of your estimates.

2. Find the data for 'Household Internet use at home by Internet activity' by clicking on Summary Tables on the left bar of the Home page and searching for 'Internet', and record them in Table 1. Compare the released data for the latest year with your estimates. Were your estimates correct?

Table 1: Average household percentage of Internet use

Year	E-mail messaging	Electronic banking	Purchasing goods and services

3. Using the information from Table 1, create line graphs for each of three purposes listed.

Exercise 2 – E-mail growth world-wide

Table 1 reveals that the percentage of households where the Internet is being used for e-mail purposes is rising. In a few years time, this percentage will hit close to the 100% mark for "regular home users" and it will not be surprising if everyone with Internet access uses the Web to send and receive e-mail messages.

1. Using your graph showing the percentage of Canadian households who use Internet access for e-mail purposes, research comparable data for other countries and include these figures in your graph.

   The following is a list of websites that might be helpful in your research:

   - National statistics offices
   - International statistical organizations

2. Where might you go to verify the accuracy of your graph? Your graph will compare how quickly Canada and other countries are becoming accustomed to using the Internet for e-mail purposes.

Exercise 3 – E-banking and e-commerce

With the results available, it is difficult to describe with certainty the percentage of households that use the Internet for the purposes of e-banking or e-commerce.

1. Under the assumption that the growth of e-banking and e-commerce is linear, find the regression equation (line of the best fit) for both sets of data. Use these equations to determine when the percentage for both types of Internet use will first surpass the 50% mark.

2. For the purpose of this question, assume that anyone who buys goods online, also uses e-banking, but not necessarily the other way around. Notice that in recent years, there has been a fairly consistent gap between those who e-bank and those who purchase goods online.

   a. Do the regression equations found in Exercise 3, part 1) indicate that this gap will remain at around 11% over the next few years? Do you think that at some point, those who e-bank will eventually begin shopping online?

b. In general, do you agree that people feel more comfortable using the Internet for electronic banking than for online shopping?

c. If they do, does this add credibility to the assumption made in Exercise 3, part 2) concerning online shoppers who also bank electronically?

3. Many banks are beginning to charge higher fees for customers who prefer to do their banking in person. At the same time, these institutions are cutting back on the number of bank tellers available.

4. How might these changes affect the percentage of people who bank or shop online?

5. What other factors or events might have a profound influence on the percentage of these Internet users?

Security measures

Security concerns are one of the main reasons why some people do not use the Internet for electronic banking or online shopping. As noted earlier, this is ironic because some of these people seem to have no reservations about sending confidential information via e-mail. There seems to be a belief that e-mail messages are more secure than online financial transactions: this is the viewpoint that certainly appears to be flawed. But this does raise an interesting question—how do organizations prevent criminals from accessing confidential information such as credit card and bank account numbers?

Businesses, banks and government organizations use numerous methods to protect their clients' personal information. Some of these methods include computer virus monitoring equipment, backup information systems and firewall (one-way access to the Internet) installation. All of these security measures are considered effective ways of blocking unauthorized access, electronic viruses and denial-of-service attacks on a computer network. Another security practice that is commonly used is data encryption. This method uses cryptography, the study of secret codes and ciphers, to hide information by creating mathematical codes.

Exercise 4 — Cryptography

Research some of the methods that are used to ensure Internet security. For example, examine how crytography uses prime numbers to keep data safe and secure.

We recommend reading a book entitled *In Code: A Mathematical Journey* (Workman Publishing, 2001). This book was written by Sarah Flannery, a 16-year-old student from Ireland, and her father, David Flannery. In it, Sarah tells the story of how she fell in love with mathematics and gives a detailed account of her research into cryptography. She also explains the theories and procedures of the new algorithm she developed for data encryption of electronic communications. For more information about *In Code* and Sarah's work, visit the following websites:

- Mathematica and the Science of Secrecy
- Early History of Cryptology

Tip! If you are interested in finding additional information about cryptolography, try searching the Internet using the keyword "RSA 129".

Exercise

1. One evening, select a Canadian television channel and watch the news. Write down the subject of any news story that uses statistical information. Were any of the decisions reported in the stories based on statistics?

2. Study three national newspapers with the same date. Count the number of articles on the front page of each newspaper. How many stories on each front page mention statistics? Calculate the ratio of articles that use statistics to those that do not. For each newspaper, calculate the percentage of stories on the front page that use statistics. Present the results in the form of a table.

3. Imagine you are a politician who wants to lower the voting age of the population from 18 to 16. What statistical information might you use to support your case? Would you argue that the decision in favour of lowering the voting age should be based on the statistics alone? Why or why not?

4. Other than statistics, what kinds of information would you need to make a decision about the following issues?
    - increasing taxation
    - reducing traffic congestion
    - quitting smoking
    - buying a computer
    - starting a new business.

5. Which city outside Canada is your favourite? Use one item of statistical information to argue in favour of your choice.

6. Can you think of situations where the same statistical information could be used to justify opposite decisions?

7. Write a one-page essay describing three uses of statistical information in society. Give examples.

8. Carefully study the world population figures in Table 1 and answer the questions below:
    - Which organizations might want these figures?
    - For what issues would these figures be relevant?
    - Based on these issues, what kinds of decisions would you make considering the information in the table below? Discuss in class.

Table 1. United Nations world population forecast

	1996	2050
	millions	
China	1,232	1,517
India	945	1,533
Pakistan	140	357
Nigeria	115	338
Indonesia	200	318
Iran	70	170
USA	269	347
Ethiopia	58	213
Brazil	161	243
Bangladesh	120	218
Kenya	28	66
Brazil	93	154
Russian Federation	148	114
Philippines	69	130
Uganda	20	66

Source: United Nations Department for Economic and Social Information and Policy Analysis. World Population. 1996.

Problems with using information

> Statistical thinking will one day be as necessary for efficient citizenship as the ability to read and write—*H.G. Wells*

The above quotation was recorded at the beginning of the 20th century, and few would disagree with its relevance today. Because the decisions that affect them are often based on some form of statistical information, Canadians need to be aware of the potential problems with gathering, understanding and presenting statistical data. No doubt the misuse of statistics, as well as other forms of information, factors heavily in the public's skepticism of the validity of some information.

This chapter will outline some of the problems you may encounter if you are not careful when using statistics.

The efficient and accurate use of information is handicapped by defects within the existing statistical structure. Sometimes, the required data do not exist. Similarly, data from different sources are not always comparable. Also, the quality of some information may need improvement. For these reasons, some uses of data may be limited or discouraged altogether.

There are many types of users who deal with statistical information in a number of different ways. With some foresight at the planning stage, the data can be made accurate, relevant and useful to an assortment of different users.

- First, it is important to determine the end use for the data. For example, suppose that your municipality wants to evaluate where a new highway should be constructed. The highway planners would have to consider or gather several types of information. The planners may conduct environmental impact surveys or gather statistical information on average land costs (to anticipate expropriation costs), demographic patterns, construction and labour cost forecasts, and the number of local residents. Although this statistical information belongs to different sub-categories, the planners must ensure that the data are accurate and appropriate, and that they serve their ultimate purpose—finding the optimum location for the new highway.
- Second, the timeliness of the data is also relevant. The type of data determines how frequently they are collected. Our national financial measures, such as gross domestic product, are collected four times a year, in quarters (each year is divided into four three-month segments). In contrast, the census—a comprehensive assessment of demographic information—is performed every five years because of its cost, scope and long-range view. Since data availability affects the reference period of a survey, then this factor must also be taken into consideration.
- Third, classifying data users by the nature of their businesses is a useful exercise. This helps the statistician and the analyst to determine what kind of information is needed. Knowing who the users are, however, does not itself shed much light on the purposes or the manner in which the data will be used.
- Finally, it is important for the user to understand the survey methodology, including the concepts and definitions. This will allow for a better understanding of the scope of the survey and how to use the data in an appropriate manner.

Misinterpretation of statistics

Misinterpretation is a common problem when using statistical information. It may be caused by a number of factors.

This section explains how statistics can be misused by

- Misunderstanding the data
- Using incomparable definitions
- Deliberately misinterpreting the information

Each day, we are bombarded by numbers in the media. Many of the facts and figures quoted in the news, such as unemployment and divorce rates, originate from *The Daily*, Statistics Canada's first and official release of statistical data and publications produced by Statistics Canada. It presents analysis of newly released data with source information for more detailed facts. Although the information from *The Daily* is objective, it is interesting to analyse how newspapers from different regions can put a different spin on the same facts.

Misunderstanding the data

Sometimes, data are misunderstood by the media. Example 1 shows how crime patterns can be oversimplified and misinterpreted.

Example 1 – Crime statistics

The Daily published data on crime statistics with the following subheadings in the release:

The Daily
Statistics Canada

The Daily, Tuesday, July 18, 2000

Crime statistics

Crime statistics, 1999

- Violent crime declines in all categories
- Large decline in break-ins
- Drug offences on rise, impaired driving stable
- Youth crime down for seventh consecutive year

Based on the information in the release from *The Daily*, Canadian newspapers headlines read:

- **Hot economy stalls murder rate**
- **Declining homicide rates knock steam out of election platforms**
- **Homicide rate at a 32-year low**

The newspaper headlines above strayed from the content of the main feature. Focusing on one aspect of the data, the newspapers ignored the main finding—that the national crime rate had declined for the eighth consecutive year. Secondly, the data revealed that the rate for violent crime fell 2.4% in 1999, the seventh consecutive decrease after 15 years of increases. All major categories of violent crime declined in 1999, including homicide (-4.7%), attempted murder (-8.8%), assault (-2.0%), sexual assault (-7.3%), and robbery (-1.5%).

The release also reported that the national homicide rate has generally been falling since the mid-1970s. That trend continued in 1999; 536 homicides were reported by police, 22 fewer than the previous year. The 1999 homicide rate—1.76 homicides for every 100,000 people—was the lowest since 1967.

Although it is possible to speculate that rising prosperity may have contributed to a lower crime rate, the data provided by Statistics Canada did not measure the relationship between the economy and crime statistics. And if it had, the observation would appear to be that the crime rate had been decreasing since 1991, the year that the Canadian recession began.

The Daily did mention that numerous factors contribute to changes in the crime rate. However, it looks as if some journalists took this statement to be a fact. The connections found in the newspaper headlines are purely speculation and were not revealed in the objective data. The journalists misinterpreted the release most likely because they misunderstood the underlying causes and effects of crime.

Statistics Canada representatives spend much time reviewing the media use of release data each day. These representatives also answer media questions regarding the data and make certain that the data are properly understood. If a misunderstanding has occurred, then the representatives try their best to correct it.

Using incomparable definitions

It is important to understand the statistical definitions and concepts behind the information that you are using. If you are examining labour force issues, you should become familiar with the definitions for terms such as *unemployment*, *employment*, and *participation rate*. If you are looking at data on environmental issues, you will need to consider the definition and concepts associated with words such as *forest*, *woodland*, *extinct*, *endangered species*, and *national park*.

A great advantage of statistical information is that it can be compared, allowing trends and characteristics to be revealed. For example, one can compare the weather of Vancouver with that of Halifax, past sporting results with the present, or the academic performance of men with that of women.

However, problems can arise in the comparison of statistics when the underlying definitions, classification or methods of data collection are different. This is especially true for statistics from different sources. Nowhere is this more apparent than with vital statistics. Consider Table 1 below.

Table 1
Married people by sex, 1998 to 2002

Number of persons	Married				
	1998	1999	2000	2001	2002
Both sexes	**14,630,173**	**14,711,793**	**14,806,694**	**14,913,766**	**15,018,130**
Males	7,299,132	7,337,226	7,381,266	7,431,522	7,476,537
Females	7,331,041	7,374,567	7,425,428	7,482,244	7,541,593

Source: CANSIM table 051-0010.

According to Statistics Canada, the definition of *married* includes people who are legally married and living together, those who are legally married and separated, and those who are living in common-law unions. If you were to compare the numbers in the table with numbers from a survey that did not include 'common-law' in the 'married' category, the results would be very different since the definitions are not the same. Look at the data in the above table. Why do they indicate that there are more married women than men? Logically, the two numbers should be the same. What appears to be happening here is that individuals are attaching their own definitions to the term *married*, and this causes the numbers to be slightly different.

Deliberately misinterpreting the information

Factual information must have integrity, objectivity and accuracy. Yet it is important to recognize that information can be misinterpreted by personal bias, inaccurate statistics, and even by the addition of fictional data.

Consider this quotation below:

> Political tacticians are not in search of scholarly truth or even simple accuracy. They are looking for ammunition to use in the information wars. Data, information, and knowledge do not have to be true to blast an opponent out of the water.
> *—Alvin Toffler*

Toffler may be overly cynical in his point of view, but in reality, people and organizations do manipulate information for their own uses. For this reason, you should always be critical about the information that is provided to you. Make certain you know where the information is coming from and find out whether or not the source is credible. Also, try to find out what methodologies were used to collect and process the data.

Finally, it is important to know how accurate the statistics are because surveys are subject to two types of errors: sampling and non-sampling errors. Sampling errors occur in sample surveys because only a portion of the population is studied, not the entire population. Non-sampling errors are present in both sample surveys and censuses.

Sampling error

When undertaking any sample survey, it will be subject to what is known in statistics as sampling error.

Sampling error arises from estimating a population characteristic by looking at only one portion of the population rather than the entire population. It 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.

Characteristics

Sampling error

- generally decreases as the sample size increases (but not proportionally)
- depends on the size of the population under study
- depends on the variability of the characteristic of interest in the population
- can be accounted for and reduced by an appropriate sampling plan
- can be measured and controlled in probability sample surveys.

Sample size

As a general rule, the more people being surveyed (sample size), the smaller the sampling error will be. Many people are surprised by the small size of well-known surveys. For example, polls that try to predict voting patterns are taken from sample sizes ranging from 1,000 to 2,000 people, with samples of about 1,000 people being the most common. Ratings for television programs are estimated from approximately 2,000 viewers. This small sample represents the television preferences of a total population of 12 million Canadian households! Despite a widely-held perception that such polls are reliable, some statisticians question their accuracy because of the small sample size.

If one of the survey objectives is to look at sub-populations or measure rare events, then a larger sample will be needed. However, it is important to note that increasing the sample size also means increasing costs.

Population size

Except for very small populations where the relationship is more direct, the size of a sample does not increase in proportion to the size of the population. In fact, the population size plays an almost non-existent role as far as large populations are concerned.

Variability of the characteristic of interest

In general, the greater the difference between the population units, the larger the sample size required to achieve a specific level of reliability. For example, if you were to conduct a survey on work environments for a population where the income varies from $30,000 to $50,000, you would use a smaller sample size to achieve the same level of reliability than you would use for a population of equal size for which income varies from $5,000 to $1,000,000.

Sampling plan

It is important to develop an efficient sampling plan, which includes a sample design and an estimation procedure. The method of sampling, called "sample design", can greatly affect the size of the sampling error. Many surveys involve a complex sample design that often leads to more sampling error than a simple random sample design. The estimation procedure also has a major impact on the sampling error. (These concepts are examined in greater detail in the chapter entitled Sampling methods.)

Measuring sampling errors

There are methods that estimate sampling error for probability sample surveys. The sampling variance is the most commonly used measure to quantify sampling error, and like the other methods, it is derived directly from the sampling and estimation methods used in the survey. (Sampling variance is examined in more detail in the chapter entitled Sampling methods.)

Non-sampling error

Aside from the sampling error associated with the process of selecting a sample, a survey is subject to a wide variety of errors. These errors are commonly referred to as "non-sampling errors".

Non-sampling errors can be defined as errors arising during the course of all survey activities other than sampling. Unlike sampling errors, they can be present in both sample surveys and censuses.

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

- **Random errors** are the unpredictable errors resulting from estimation. They are generally cancelled out if a large enough sample is used. However, when these errors do take effect, they often lead to an increased variability in the characteristic of interest (i.e., the greater the difference between the population units, the larger the sample size required to achieve a specific level of reliability).
- **Systematic errors** are those errors that tend to accumulate over the entire sample. For example, if there is an error in the questionnaire design, this could cause problems with the respondent's answers, which in turn, can create processing errors, etc. These types of errors often lead to a bias in the final results.

Non-sampling errors are extremely difficult, if not impossible, to measure. Since random errors have the tendency to be cancelled out, systematic errors are the principal cause for concern. Unlike sampling variance, bias caused by systematic errors cannot be reduced by increasing the sample size.

Characteristics

Non-sampling errors

- can occur in all aspects of the survey process other than sampling
- exist in both sample surveys and censuses
- are difficult to measure

Non-sampling errors can occur because of problems in coverage, response, non-response, data processing, estimation and analysis. Each of these types of errors is explained below.

Coverage errors

An error in coverage occurs when units are omitted, duplicated or wrongly included in the population or sample. Omissions are referred to as "undercoverage", while duplication and wrongful inclusions are called "overcoverage". Coverage errors are caused by defects in the survey frame, such as inaccuracy, incompleteness, duplications, inadequacy or obsolescence. Coverage errors may also occur in field procedures (e.g., while a survey is conducted, the interviewer misses several households or persons).

Response errors

Response errors result when data is 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.

  For more information, refer to the section on Questionnaire design.
- **Interview 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 errors**
  Respondents can also provide incorrect answers. Faulty recollections, tendencies to exaggerate or underplay events, and inclinations to give answers that appear more 'socially desirable' 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 (taking answers from someone other than the respondent) or lacking control over the survey procedures are just a few ways of increasing the possibility for response errors.

Non-response errors

Non-response errors are the result of not having obtained sufficient answers to survey questions. There are two types of non-response errors: complete and partial.

- **Complete non-response errors**
  These errors occur when the results fail to include the responses of certain units in the selected sample. Reasons for this type of error may be that the respondent is unavailable or temporarily absent, the respondent is unable or refuses to participate in the survey, or the dwelling is vacant. If a significant number of people 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 non-response errors**
  This type of error occurs when respondent provide incomplete information. For certain people, some questions may be difficult to understand. To reduce this form of bias, care should be taken in designing and testing questionnaires. Appropriate edit and imputation strategies will also help minimize this bias.

More information on editing and imputation can be found in the chapter entitled Data processing.

Processing errors

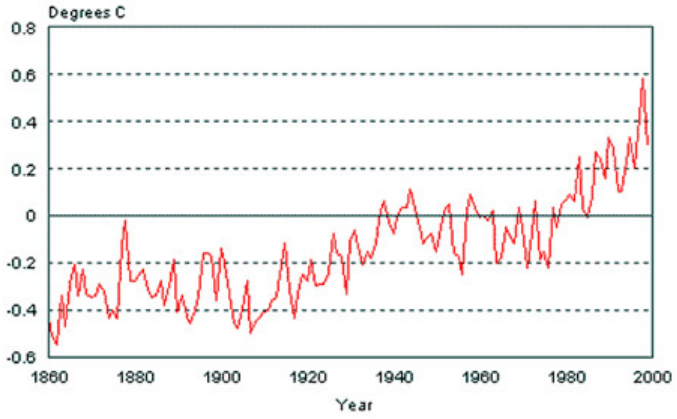
Processing errors sometimes emerge during the preparation of the final data files. For example, errors can occur while data are being coded, captured, edited or imputed. Coder bias is usually a result of poor training or incomplete instructions, variance in coder performance (i.e., tiredness, illness), data entry errors, or machine malfunction (some processing errors are caused by errors in the computer programs). The same thing can be said about captured errors. Sometimes, errors are incorrectly identified during the editing phase. Even when errors are discovered, they can be corrected improperly because of poor imputation procedures.

Estimation errors

Statistics Canada and other data-collecting agencies devote much effort to designing and monitoring surveys in order to make them as error-free as possible. If an inappropriate estimation method is used, then bias can still be introduced, regardless of how errorless the survey had been before estimation.

Here is an example of a potentially inappropriate estimation. We know that global warming is an issue where there is a lot of debate. To accurately measure this phenomenon, one should know how to come up with an acceptable "average global temperature". Figure 1 features a common portrayal of climate change data. It shows an average global temperature increase between 0.3° and 0.6°C over nearly 140 years.

Figure 1. Global climate change, 1860 to 2000



The measurements that comprise the data set have been taken at various weather stations around the world. In this case, the population is the set of weather measurements, from which a sample can be taken.

Some scientists question the accuracy of a graph like Figure 1 because they feel that the estimates from the sample survey are biased.

Scientists argue that measurements of temperature should reflect the ratio of the earth's land mass to the water mass. For example, if the land mass is half of the mass of water (seas and oceans), then twice as many measurements should come from locations over water than over land. In fact, in Figure 1, few measurements were taken from locations over the surface of water, whereas the great majority of measurements were taken from weather stations on land.

Why might this bias the estimates from the sample survey?

Temperatures on land tend to be naturally higher than on water surfaces owing to the phenomenon known as 'urban heat island effect.' If the sample is too heavily weighted in favour of land-based temperatures, and the estimates do not take this into account (as some scientists claim), then the results may not truly reflect a global average.

For more information on estimation, refer to the Sampling methods chapter.

Analysis errors

Analysis errors are those that occur when using the wrong analytical tools or when the preliminary results are provided instead of the final ones. Errors that occur during the publication of data results are also considered analysis errors.

Summary

Whenever you are presented with statistical information, it is useful to have a list of questions ready to help you judge the reliability of the statistics. This is not to suggest that there are problems with the data, but rather to help you gain confidence in questioning the reliability of the data.

The following are just some of the questions you should ask when presented with statistical information. Remember that if the source cannot provide you with answers or explanations, then you should question how sound the data really are.

- Who is the author (source) of the information?Is the source primary (i.e., the organization that collected the data) or secondary (an outside analyst or organization)?
- Does the primary source of information have a reason for misrepresenting the information?
- If the information is derived from a secondary source, is it possible that the data might have been altered for any reason?
- Is it necessary to find out the method of data collection, sampling technique or response rate to the survey?
- If the information is taken from a sample survey, do you think the sample size was adequate? What is the level of sampling error?
- Were the survey questions easy to understand?
- Do you understand the definitions of variables or topics discussed in the survey or census?
- Are the definitions consistent?

The link *Finding and Using Statistics* will tell you how to find information that is reliable and what to do with this information. This link covers how to find the appropriate data; how to begin; how to read statistical tables; symbols and definitions; putting data in context; use of basic techniques; and drawing conclusions.

Exercise

1. Determine which of the following statements are reliable. Using your knowledge from this chapter, list the possible problems with each statement.
    a. The average annual income of Canadians is $29,100 according to a survey carried out in Oshawa.
    b. A large majority of people from rural areas support subsidies for failing farm operations. This is the result of a phone-in poll carried out by a regional television station.
    c. Statistics reveal that 30% of our nation's school leavers are below average in reading and writing.
    d. Canada's youth unemployment rate is over 12%; therefore, more than 12% of Canada's 15- to 24-year-olds are unemployed.
    e. Recently, a leading environmental group claimed that only 3% of Canada's land mass was forest cover, whereas a leading business organization claimed the figure was 6%.
2. What are the two types of non-sampling error?
3. What is sampling error influenced by?
4. List the possible sources of systematic (or bias) errors.

Class activity

1. Conduct a class debate on the following idea:
    - Expressing statistical information only through tables and graphs is not enough—data need to be supported by text.

Answers

1.

   a. No, this statement is not reliable. This statistic is a result of an inappropriate estimation because of its unrepresentative sample. Choosing Oshawa as a representative population neglects regional variations such as type of industry (e.g. resource-based, tourism, manufacturing, high-tech) or employment (e.g. seasonal, permanent). In order to have a representative sample, the statistician must gather information from many different areas across the country, including rural, semi-urban and urban areas.

   b. No, this statement is not reliable. Participants in a TV station phone-in poll select themselves voluntarily, which can lead to an unrepresentative sample.

   c. This statement may or may not be reliable. Its reliability is questioned because we do not know the source of the statistics about school leavers.

   d. Again, this statement may or may not be reliable. We do not know the origin of this information. However, if the survey was conducted properly nationwide, and if 'youth' is defined as those people between the ages of 15 and 24 years, then there is no reason to state that this statistic is incorrect.

   e. No, this statement is not reliable. If the statistics are conflicting, this could mean that different definitions of "forest cover" are used or that one of the sources may be flawed, or that the results have been misinterpreted by one of the parties.

2. The two types of non-sampling errors are random errors and systematic (or bias) errors.
3. Sampling error is influenced by the size of the sample and of the population, the variability of the characteristic of interest in the population, the sample design and the estimation method.
4. Systematic errors can result from
   - coverage errors
   - response errors
   - non-response errors
   - processing errors
   - estimation errors
   - analysis errors.

Confidentiality, privacy and security

As society comes to depend more and more on information, new problems of individual rights and privacy arise. People want information about many things, such as the latest figures on jobs for school leavers and up-to-the-minute accurate bank balances from automatic teller machines. At the same time, many people are concerned that too much of their information is being stored on computer databases and accessed by persons or organizations unknown to them.

Writers and filmmakers have painted terrifying pictures of futuristic societies in which the thoughts and actions of human beings are controlled by all-powerful computers. Such fears are often exaggerated in the interests of good fiction. However, there are important issues that society must handle to ensure that the rights of the individual are protected in the information age. This chapter discusses some of these issues.

Providing information

People provide information about themselves all the time. To rent a video or borrow a library book, it is necessary to complete a personal information form. Taxpayers and users of government services must also provide details about themselves.

Likewise, banks and large retailers will only issue credit cards to customers if they know something about their income, occupation and family status. Health services collect and store data about each patient they treat.

These are just a few examples of where people are required to provide personal information, much of which is entered into computer databases. The concern is that authorities and other organizations may be able to access and link to such databases. In this way, data profiles (or information pictures) of individuals could be put together and perhaps used in a way that compromises the individual.

In order to protect privacy, personal information must not be used for purposes other than those for which its collection was authorized.

There is also concern that individual privacy and corporate confidentiality may be breached by someone's hacking into computer databases. Hacking usually applies to computer users who gain unauthorized access to large databases and may even amend the data files. Such activity is highly illegal.

Concerns about information privacy are legitimate, but there are many beneficial aspects to the provision of information. Modern society needs information in order to function efficiently. By the same token, people who provide information need to be sure that the confidentiality and security of their data are protected.

Privacy and security

Increasingly, people need more information and better skills in handling it in order to make decisions. As the information age evolves, privacy remains an essential issue to be considered.

In Canada, government and community organizations are always trying to find ways to improve the confidentiality and security of statistical information. For example,

- holders of data are encouraged to maintain tight security procedures so that only those with authorization can access databases;
- much more secure telecommunications links are now available for transmission of data between different locations of a particular government agency or business organization; and
- tighter procedures are also in place to counter the spread of computer viruses.

Statistics Canada's values

> *Reliability, objectivity and confidentiality are essential, and mutually supportive, in the functioning of the profession of statistics and the operation of a statistical agency.*
> *— Dr. Martin B. Wilk, Chief Statistician 1980-1985*

For Statistics Canada, safeguarding the privacy and security of the information provided to the Agency, is a concern of the utmost importance. To make certain that this concern is taken into consideration at all times, we have developed a series of values to help us achieve our goal. These values include:

- ensuring objectivity;
- protecting confidentiality;
- focusing on analysis;
- reducing the response burden; and
- establishing professionalism and reliability.

Privacy and security at Statistics Canada

As the national statistical agency, Statistics Canada takes strong measures to ensure that the confidentiality and security of data provided by individuals, businesses and organizations are carefully protected.

All Statistics Canada employees take an oath of secrecy and face severe penalties for any breach of confidentiality. Employees who break the oath may be fined and/or jailed for up to six months.

Information collected under the *Statistics Act* cannot be disclosed under the *Access to Information Act* or any other act. The Canada Customs and Revenue Agency, the Royal Canadian Mounted Police and the courts do not have access to survey responses.

Access ID strictly controlled

All Statistics Canada employees are responsible for ensuring the security of confidential information. Only employees who need to view confidential files as part of their duties are authorized to access them. A network of physical security systems and procedures protects confidential information against unauthorized access.

An important element of Statistics Canada's security system is the electronic protection of survey data stored in computer databases. Statistics Canada uses two completely different computer networks, one internal and one external. Confidential data are stored, transferred and processed on the internal network only, which has no connection with any external network. This prevents any possible external access.

Precautions against disclosure

Statistics Canada publishes data as statistical summaries, tables and graphs. No data that could identify an individual, business or organization, are published without the knowledge or consent of the individual, business or organization. All final results are carefully screened before release. For example, if only two companies make a particular product in one province, we will not publish the provincial sales figure for that product without the consent of both companies. The same is true if three companies make the product, but one or two of them have an overwhelming share of the provincial market. If the figure for one province is not published, the figure for at least one other province is suppressed, to avoid the possibility of working out the missing figure from the national total.

In addition to this suppression approach, other methods can be used in order to eliminate potential risk of disclosure. Some of these methods include: rounding, collapsing, micro-aggregation and data swapping.

What Statistics Canada is allowed to disclose

While the *Statistics Act* ensures the confidentiality of survey responses, it also provides a way to avoid burdening the respondent with duplicate surveys. The Act allows the Agency to enter into sharing agreements with federal and provincial departments and ministries and with corporations. Respondents are informed at the time of collection if a data-sharing agreement applies to their particular survey.

Respondents may also permit the disclosure of their information by signing a waiver which grants permission allowing a specific topic to be released. Releases of identifiable information are governed by a policy and each must be authorized by the Chief Statistician of Canada.

Exercises

1. Make a list of some organizations where personal data about you might be stored.
    a. List some of the uses that might be made of the data.
    b. Which uses do you think are appropriate?
    c. Are there any that you consider inappropriate? If so, why?

2. Does your school record your personal data in a computer? If so, what sort of data? What type of information is available for you to look at? Is any of the information private? If so, why?

3. Imagine you are in charge of computer security in a government agency or business. What would you do to ensure that personal information held in the organization's databases was not accessed without the proper authority?

Computers and data

Professional organizations, institutions and businesses have been using computers to process data and provide information for many years. Over time, computers and computer systems have been growing in both sophistication and complexity, but their basic characteristics remain the same.

Simply put, computer systems combine several ingredients: *hardware*, *software* and *users* (people). Each is necessary for the system to function.

The electronic computer was introduced to the Canadian business community and the world in the 1960s. Its adoption into society was gradual because early models were expensive to purchase and maintain, slow to perform, and very little expertise was available. Twenty-five years ago, only about 10% of staff and students at a typical university were using computers—today the figure is close to 100%. Furthermore, commercial computing skills were not taught at university, and business software applications could not be bought—they had to be developed in-house by programming staff. However, skilled programmers were scarce and expensive.

We often refer to the 'generations' of computing history.

- First generation computers (1940 to 1950) were based mostly on wired circuits of vacuum tubes and punched cards were the main storage medium.
- Second generation computers (1950 to 1964) used transistors. The inefficient vacuum tube was replaced with a much smaller and more reliable component.
- Third generation computers (1964 to 1972) used integrated circuits. This invention led to the widespread use of computers today. Scientists found a way to reduce the size of transistors, enabling computer manufacturers to build smaller and less costly computers.
- Fourth generation computers (1972 to present) use microprocessor chips, which are large-scale integrated circuits containing thousands of transistors. The transistors on a microprocessor chip are capable of performing all of the functions of a computer's central processing unit. The microprocessor led to the creation of the first personal computer.

History of computers

The history of computers, and the history of computers in education

500 BC

Evidence of the abacus, the world's first calculating machine, exists from as far back as 2,500 years ago in the Tigris-Euphrates Valley. The earliest form of the abacus is a stone or clay tablet that uses pebbles for counting. Grooves are carved into the tablet and pebbles, representing numbers, are placed in these grooves. The pebbles can slide along the grooves from one side of the tablet to the other, thus allowing for easier counting. The abacus also helps ancient peoples perform simple calculations such as addition and subtraction.

1300 AD

In 13th century China, the idea was to thread beads or drilled-out pebbles onto string or wires attached to a wooden frame. This becomes the basis for the modern-day abacus.

Early 1600s

John Napier, the Scottish inventor of logarithms and the decimal point, invents a hand-held device to help with multiplication and division. His device is known as *Napier's Rods* or *Napier's Bones*.

1622

William Oughtred of England invents the *Slide Rule*. Unlike the slide rules of the future, his is circular in shape.

1623

Wilhelm Schickard of Germany makes a calculating machine called the *Calculating Clock*. It is capable of adding and subtracting up to six digits. The machine and its plans were lost but were rediscovered in 1935, lost again and found once more in 1956. In 1960, Schickard's machine is later reconstructed and found to have worked.

1642

Blaise Pascal of France builds the *Pascaline*—the first digital computer that can add figures. Up to eight digits can be entered into the machine by turning dials. Pascal actually built and sold about a dozen of these, and some of them still exist today.

1668

Sir Samuel Morland of England produces a non-decimal adding machine suitable for use with English money.

1673

Gottfried Wilhelm Leibniz of Germany, the co-inventor of calculus, designs the *Stepped Reckoner*, a machine that can carry out the multiplication of up to 12 digits. It is also capable of dividing and finding square roots as well as adding and subtracting. The machine was lost in an attic until 1879.

1775

Charles, the third Earl of Stanhope of England creates a successful multiplying calculator.

1801

Joseph-Marie Jacuard develops an automatic loom controlled by punch cards.

1820

Charles Xavier Thomas de Colmar of France develops the *Arithmometer*, the first mass-produced calculator that can successfully add, subtract, multiply and divide numbers.

1822

Charles Babbage of England designs the first mechanical computer.

1832

Babbage produces a prototype for the first automatic mechanical calculator. The function of his *Difference Engine* is to calculate and print mathematical tables. Only one-seventh of the engine is ever assembled by Babbage's engineer, Joseph Clement. Most of the 12,000 manufactured parts are later melted for scrap.

1833

Babbage begins designing the *Analytical Engine*. This machine has storage systems and computing components such as input and output units.

1853

To produce a set of astronomical tables, the Dudley Observatory in Albany, New York buys the first tabulating machine built by Swedes George Scheutz and his son Edvard. Their machine is based on Babbage's design.

1854

British mathematician George Boole devises binary algebra. His work is the basis for binary switching, upon which modern computing depends.

1878

Ramon Verea, a Spaniard living in New York, invents a calculator with an internal multiplication table.

1886

Herman Hollerith of the United States Census Bureau develops a mechanical device that uses punched cards to compile and tabulate data. Dorr. E. Felt of Chicago constructs the first calculator to enter numbers by pressing keys instead of turning dials. His calculator is called the *Comptometer*.

1889

Felt invents the first printing desk calculator.

1896

Hollerith establishes the Tabulating Machine Company which eventually becomes the International Business Machines (IBM) corporation.

1931

E. Wynn-Williams uses a thyratron tube to construct a binary digital counter for use in physics experiments at Cambridge University.

1935

IBM introduces the *IBM 601*, a punch card machine with an arithmetic unit based on relays that is capable of doing a multiplication per second.

1937

George Stibitz of Bell Telephone Laboratories constructs a one-bit binary adder using relays. This is one of the first binary computers. Alan Turing develops his *Universal Machine*.s

1939

John Vincent Atanasoff and Clifford Berry of Iowa State College completes a prototype 16-bit adder. This is the first machine to calculate using vacuum tubes. Bell Telephone Laboratories develops the *Complex Number Calculator*.

1941

Atanasoff and Berry completes a special-purpose calculator for solving problems of simultaneous linear equations. It is later called the *Atanasoff-Berry Computer* (*ABC*). This computer has 60 fifty-bit words of memory in the form of capacitors mounted on two revolving drums. Its second memory burns holes into punch cards.

1943

Thomas Flowers of England builds the earliest programmable electronic computer which contains 2,400 vacuum tubes. Called the *Colossus Mark I* decrypting computer, it translates 5,000 characters per second and uses punched tape for its input. This computer is developed to crack the German coding device, *Enigma*.

1945

A bug is found in a computer relay, and the term "debugging" is coined.

1946

The first vacuum tube-based computers are developed at the University of Pennsylvania. The *Electrical Numerical Integrator and Calculator* (*ENIAC1*) has 18,000 vacuum tubes and takes up 1,800 square feet of space. It is considered the first "true computer" (i.e., the first fully electronic, general purpose digital computer).

1947

The transistor is invented by William B. Shockley, John Bardeen and Walter H. Brattain at the Bell Telephone Laboratories in the United States.

1951

The baby boom causes an increase to classroom size, but little electronic technology is used in schools. The first generation *Universal Automatic Computer* (*UNIVAC*) computer is delivered to the United States Census Bureau. *Whirlwind*, the first real-time computer is built for the U.S. Air Defense System.

1952

UNIVAC is used to predict the 1952 United States presidential election. No one believes its prediction, based on 1% of the vote, that Eisenhower will sweep the election. He does.

1955

IBM sells its first commercial computer.

1957

The first transistorized computer, the *Transistorized Experimental Computer* (*TX-O*) is completed at Massachusetts Institute of Technology.

1958

Mainframe host computers are not widely accepted in schools, which are still using the single classroom, teacher-as-manager method of delivering information to students. Jack St. Claire Kilby invents the integrated circuit at Texas Instruments. The late 1950s sees the development of two important computer-programming languages—*Common Business Oriented Language* (*COBOL*) and *List Processor* (*LISP*).

1959

The transistor-based computers comes into use. Smaller computers based on transistors and printed circuits are built between 1959 and 1964. These are regarded as "second generation" computers.

1960

Dr. Grace Murray Hopper, professor of mathematics, finishes creating the *COBOL* language.

1962

Airlines begin to use a computerized reservation system.

1963

The U.S. *Vocational Education Act* provides new money to support technology in schools. However, the mainframes and minicomputers use batch-processing methods that do not fit well with the single teacher-as-manager-of-learning methods used in most schools. *Beginners All-Purpose Symbolic Instruction Code* (*BASIC*), a simple high-level programming language is developed and used mostly in universities to train programmers. The *IBM 360* family of computers is developed. Most computers are still using host methods with punch cards as the primary input device. Line-printers are still the primary output device. The first microcomputer, called the *PDP-9*, is built by Digital Equipment.

1964

Computers built between 1964 and 1971 are regarded as "third generation" computers; they are based on the first integrated circuits, which enable machines to be made even smaller. IBM releases the *PL/1* programming language. The *IBM 360* is launched. The computer mouse and "windows" are invented this year, as well as the programming language *BASIC*. Gordon Moore (Chief Executive Officer or CEO of Intel between 1979 to 1987) predicts that the number of transistors the industry would be able to place on a computer chip will double each year. This theory later becomes known as the Moore's Law.

1965

Mainframes and minicomputers are put into place in some schools. These are used mainly for administrative purposes and school counselling. The first supercomputer, the *Control Data CD6600* is developed.

1967

High-level programming languages such as *Formula Translation* (*FORTRAN*) are being taught in universities. School vocational training programs begin to include computer maintenance.

1968

The Intel Corporation is founded.

1969

The United States Department of Defense starts the Advanced Research Projects Agency network (ARPAnet) for research into networking. It is the basis for what is now the Internet. The original Net is a small network of supercomputers that is used to promote the sharing of research among universities. The first hosts are connected in 1969.

1970

Mainframes and minicomputers are used in some schools, but not extensively in the delivery of instruction. Intel introduces the first random-access memory (RAM) chip, the *1103*, with a capacity of one kilobyte or 1024 bytes.

1971

Intel's first microprocessor, the *4004*, is developed. It is capable of approximately 60,000 interactions per second (0.06 millions of instructions per second or MIPs), running at a clock rate of 108 KHz. The first microcomputers (PCs) are developed. Mainframes and minicomputers are in wide use in business. A few new software companies develop mainframe- and minicomputer-based instructional programs. The floppy disk is invented by IBM engineers led by Alan Shugart. The development of the programming language *PASCAL* (named after the famous mathematician) is completed.

1972

Computers built after 1972, referred to as the "fourth generation" computers, are based on large scale integration circuits (i.e., microprocessors) with 500 or more components on a chip. The *C* programming language is developed. Later developments included *C++*. The first hand-held scientific calculator (the *HP-35*) made by Hewlett-Packard, makes slide rules obsolete. Intel releases the *8008 Processor*. Canada's Automatic Electronic Systems introduces the world's first programmable word processor with a video screen. This computer, named the *AES 90*, uses magnetic disks for storage and custom-built microprocessors.

1974

The *Apple I* computer is sold in kit form. The *CLIP-4*, the first computer with a parallel architecture, is developed. The *MITS Altair 880*, the *Scelbi* and the *Mark-8* are introduced. These are considered the first personal computers. Telenet opens and the first commercial version of the ARPAnet is released.

1975

Some *Apple I* PCs are donated to schools. Some schools adopt mainframes and minicomputers but most refuse to consider PCs. Bill Gates and Paul Allen implement *BASIC* for the first time. IBM introduces the first laser printer. Micro-Soft was formed by Bill Gates and Paul Allen. (The hyphen in "Micro-Soft" is dropped later on.)

1977

Apple II is released.

1979

Fifteen million PCs are estimated to be in world-wide use. The PC-based spreadsheets are developed. Mainframes and minicomputers are still in wide use. The release of the arcade video game, *Space Invaders*, starts the video game craze. Honeywell introduced the programming language *Ada*, named after Augusta Ada Byron, one of the first computer scientists in history and, surprisingly, the daughter of Lord Byron, the famous Romantic poet.

1980

The *TI-99*, from Texas Instruments, uses a television screen as a monitor, the world's most popular PC. Development of *MS-DOS/PC-DOS* begins. The *Sinclair ZX80* is released.

1981

IBM develops and introduces a PC, the first mainframe manufacturer to do so. The first educational drill and practice programs are developed for personal computers. The *Xerox Start System* and the first *Windows, Icons, Menus and Pointing Devices* (*WIMP*) system are developed. ARPAnet has 213 hosts and is growing rapidly. Microsoft introduces M*S-DOS version 1.0*.

1982

The *TCP/IP* protocol is established, and the term "Internet" is used for the first time to describe the connected set of networks using *TCP/IP*. The *Commodore 64* is released. Compaq releases their *Compaq Portable*, which is IBM PC-compatible. IBM launches double-sided 320K floppy disk drives.

1983

IBM PC-clones flourished and the Sperry Corporation becomes the second mainframe manufacturer to develop an IBM PC-compatible computer (developed by Mitsubishi in Japan). The *Apple II* finds widespread acceptance in education because it fits the teacher-as-manager model of instruction. Simple simulation programs are developed for personal computers. IBM releases the *PC junior*.

1984

There are still relatively few computers in the classroom. The *Apple Macintosh* computer is developed and released. Commercial software manufacturers develop computer-based tutorials and learning games. The domain name server (DNS) is introduced to the Internet, which consists of about 1,000 hosts. William Gibson coins the term "cyberspace" in his novel *Neuromancer*.

1985

Microsoft Windows is launched. Lotus, Intel and Microsoft introduces *Lotus, Intel and Microsoft Expanded Memory Specification Standard* (*LIM EMS Standard*).

1987

The number of Internet hosts exceeds 10,000.

1988

Laptops are developed. The first optical chip is developed. *Write Once Read Many times* (*WORM*) disks are marketed for the first time by IBM.

1989

The "World Wide Web", invented by Tim Berners-Lee, sees the need for a global information exchange. The Sound Blaster card is released.

1990

Multimedia PCs are developed. Schools are using videodiscs. Object-oriented multimedia authoring tools are in wide use. Simulations, educational databases and other types of computer assisted instruction programs are delivered on CD-ROM disks, many of these with animation and sound. ARPAnet is decommissioned and the number of hosts had passed 300,000.

1991

Linus Torvalds of Finland develops *Linux*, a variant of the UNIX operating system. Intel's monopoly in the Windows PC world is challenged when AMD releases its Am386 microprocessor.

1992

Schools are using *Gopher* servers to provide students with online information.

1993

Commercial providers are allowed to sell Internet connections to individuals. *Pentium* is released. The first graphics-based web browser, *Mosaic*, becomes available. The PDF (Portable Document Format) standard is introduced by Adobe. AMD releases its Am486 microprocessor to compete with Intel's 80486.

1994

Digital video, virtual reality, and 3-D systems capture the attention of many, but fewer multimedia PCs than basic business PCs are sold. Object-oriented authoring systems such as *HyperCard*, *Hyperstudio*, and *Authorware* grow in popularity in schools. *Netscape 1.0 i*s written as an alternate browser to the *National Center for Supercomputing Applications* (*NCSA*) *Mosaic*. First wireless technology standard (Bluetooth). Yahoo! Internet search service launched. The World Wide Web comprises at least 2,000 Web servers.

1995

The Internet and the World Wide Web begins to catch on as businesses, schools, and individuals create web pages. Most of the computer assisted instruction is delivered via CD-ROM disks, which are growing in popularity. *Windows 95* is released, as well as *Pentium Pro*. Netscape announces *JavaScript*. George Moore revises his law to say that the number of transistors placed on an integrated circuits will now double every two years. Intel's supremacy in the Windows PC world continues to erode as AMD releases its K-5 microprocessor, which offers a cheaper, pin-compatible alternative to the Pentium microprocessor.

1996

The Internet is widely discussed as businesses begin to provide services and advertising using web pages. New graphics and multimedia tools are developed for the delivery of information and instruction using the Internet. Many schools are rewiring for Internet access. A few schools install web servers and provided faculty with a way to create instructional web pages. *Netscape Navigator 2.0* is released. The number of computer hosts connected to the Internet approached 10,000,000. Microsoft releases the first version of Internet Explorer, its proprietary Web browser.

1997-1998

The growth of the Internet continues to expand with new uses and applications. Voice recognition is slowly entering the computing mainstream. Educational software is expected to become more popular with the introduction of much larger CD-ROM capacities. Intel releases the *Pentium MMX* for games and multimedia enhancement. Intel releases the *Pentium II* processor. Microsoft released *Windows 98*. AMD releases the K-6 microprocessor. Palm Computing markets the first PDA (Personal Digital Assistant), the Palm Pilot. Introduction of e-Book technology. Internet-based computing starts on a large scale with downloadable programs such as SETI@Home.

1999

Linux Kernel 2.2.0 is released. The number of people running *Linux* is estimated to be about 10 million. Advanced Micro Devices (AMD) releases *K6-III*, the 400MHz version. In some tests, this computer outperformed the *Intel P-III*. It contains about 23 million transistors. Intel launches the Pentium III line of microprocessors. Many e-commerce sites are set up on the Internet. Governments and businesses all over the planet make last-minute preparations for the arrival of the year 2000 (Y2K): while new computers are fully Y2K-compliant, fears remain that there are still too many vulnerable computers in use. AMD releases its proprietary Athlon chip, which sets a new speed record of 1 GHz, outpacing all of the competing Pentium microprocessors offered by Intel.

2000

Fears of the Y2K bug prove largely groundless as the new millenium arrives without global computer network crashes. Microsoft launches Windows Millennium (for PCs) and Windows 2000 (for local area networks). Numerous Web-based businesses ("dot-coms") go out of business as their shares collapse on the stock market. Microsoft chairman Bill Gates resigns as CEO of his own company to dedicate himself to the development of software. IBM releases a follow-up to *Deep Blue,* nicknamed *Blue Gene*: it operates at 1 quadrillion ops per second (one peta flop) and is 1,000 times faster than *Deep Blue*. *Blue Gene* will be used for modelling human proteins. Cyber attacks bring down some websites.

2001

The first *Linux* virus is detected. Gordon Moore says that the law named after him should be changed again from a doubling of transistors on an integrated circuit every five years to every two years, starting between 2010 and 2020. Amazingly, considering the speed of change in this field of technology, Moore's Law has held for 36 years. Intel launches Pentium IV. Microsoft releases Windows XP (for both PCs and networks).

2002

Wireless computing becomes widespread: new handheld devices are sold which bring together wireless communications modems, dual-mode cell phones, Web browsers, palmtop computers, Global Positioning System (GPS) receivers and increasingly sophisticated operating systems and graphical user interfaces (Tablet PC).

Computers at Statistics Canada

Introduction

The *Census and Statistics Act* of 1905 ushered in the modern statistical era by creating a permanent Census and Statistics Office, known today as Statistics Canada.

Computers at Statistics Canada

1911

The Census consists of 13 questionnaires with 522 questions. The machinery used for tabulating census records from 1911 through 1942 is custom-built by A. E. Thornton, of the Bureau's mechanical tabulation staff.

1918

The Governor General in Council approves the use of mechanical appliances for all statistical compilations conducted by the government. Cards are punched at the departments themselves and then transferred to the newly created Dominion Bureau of Statistics for sorting and tabulation.

1919

The Department of Finance uses the Bureau's tabulating machines to compile income tax statistics.

1921

Compilation and tabulation are almost entirely mechanical, but productivity is greatly increased through the use of a new sorter-tabulator developed by Fernand Bélisle of the Bureau's mechanical tabulation staff. Three of these machines are built. They produce materials over 50 times as brief and concise as would have been possible with older equipment and save significant labour costs.

1932

R. H. Coats proposes the creation of a social and economic research body. He concludes that it "...would solve for some time to come, what had undoubtedly developed into one of the prime needs of government—that of keeping abreast of economic thinking, and of being continuously equipped with the materials required in the formulation of its broader economic policies, and of having more definitely recognized machinery to that end."

1949

The Mechanical Tabulation Division (MT) has 202 employees working with 173 units of tabulating equipment and 27 adding machines or comptometers. The division consists of a keypunch unit that serves several client divisions, a number of tabulating units geared to the special requirements of particular clients, and an auxiliary machine unit that supports the tabulating units. There are also calculating and adding machines and pegboards that perform simpler kinds of compilations directly from the completed statistical returns.

1950s

The U.S. Bureau of the Census used the UNIVAC computer to process the 1950 Census data. In 1951, a report stated "it would take at least 650 keypunch operators, working on 17 document punch machines during the week of peak Census processing, to produce a million punched cards completely edited and ready for tabulation." To realize productivity gains, and to pursue their interest in technical innovation, the Bureau acquired an electronic computing machine.

1960

Herbert Marshall, the Dominion Statistician, announces that the Bureau is getting ready to install an electronic computer. In 1960, an *IBM 705* computer is acquired to process the 1961 Census. It is one of the biggest computers in Canada, and the heat generated by its 10,000 vacuum tubes has to be offset by two large air conditioners.

1962-1970

In 1962, Statistics Canada has one *IBM 705*, and one *IBM 1401* with increased capacity. In 1963, an *IBM 101* is added to the list. A year later, a second *IBM 1401* is introduced. An *IBM 360/30* replaces the first *IBM 1401* in 1966. The 1966 Census is processed in the same manner as that of 1961. The major advances in these systems during this time are the *Geographically Referenced Data Storage and Retrieval System* (*GRDSR*) and the time series data bank, *Canadian Socio-economic Information Management System* (*CANSIM*). The *IBM 360/30* is upgraded to an *IBM 360/65* in 1969.

1971

The year 1971 marks the 100th anniversary of the first Census of the Dominion of Canada. The processing of population and housing questionnaires is carried out with 12 automatic tape-feed cameras provided by the United States Bureau of the Census. Data is transformed from film to magnetic tape by means of a new document reader, the *Film Optical Sensing Device for Input to Computer* (*FOSDIC*). Computer processing is carried out on an *IBM 360/65* that is substantially upgraded to double the core memory (1,024,000 bytes). A large number of tables are produced as computer printouts and then microfilmed for distribution to users. In addition, because there is a demand for aggregated data in machine-readable form, a summary tape is produced.

1972

The *IBM 360/65* is replaced by an *IBM S/370-165* with a capacity of 1,536,000 bytes.

1973

The *IBM S/370-165* now has a capacity of 2 million bytes, with an amazing 3.6 billion bytes disk storage. In addition, a mainframe, stand-alone microcomputers and minicomputers are also purchased.

1974

The mainframe capacity of the *IBM S/370-168* is 3 million bytes and the disk storage capacity is 5.2 billion bytes.

1976

Statistics Canada builds a new computer centre in the Main Building. The mainframe at this time is the *Amdahl 470/V6*, with a disk capacity of 10 billion bytes.

1981

Statistics Canada's personal computer revolution begins in a small way when a few *Xerox Star* personal computers are acquired. Then, additional IBM-compatible personal computers are acquired.

1982

Mainframe facilities are centralized, and the system is upgraded to the *Amdahl/V8*.

1985

Statistics Canada's mainframe *Amdahl* is upgraded to the *5870/5880*.

1987-1988

The mainframe computer is replaced by an *IBM 3090*, so that Statistics Canada can process 5,000 to 6,000 jobs per day—tabulating survey data or updating and querying databases. Statistics Canada purchases 200 personal computers, bringing the total to nearly 1,000 units.

1989

Statistics Canada acquires the first Automatic Cartridge Storage System which automatically stores and retrieves tape cartridges in a silo.

1990

The mainframe is replaced by an *Amdahl 5990*. In the regional offices, local area networks (LANs) are established as part of a multi-year program to upgrade Electronic Data Processing facilities.

1991

The mainframe is replaced by the *IBM 3090/600S*, which is 140 times more powerful than the system that had been purchased in 1969. The disk storage capacity is increased to 1.2 trillion bytes. There are over 4,000 personal computers—one for every employee.

1993

To improve the quality of data and reduce response burden, interviewers in regional offices test hand-held computers (laptops) in the field.

1996

The mainframe is upgraded to an *IBM 9672-R44* with 2 gigabytes of memory and fibre-optic (ESCON) channels.

1999

The availability of Unix System Services was announced giving users the choice of running traditional batch and TSO workloads or Unix workloads on the mainframe.

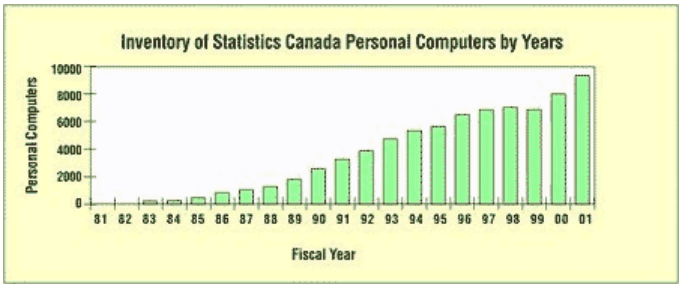
2001

The mainframe disk subsystem was upgraded to provide a capacity of 2.5 trillion bytes of RAID protected disk storage capacity.

2002

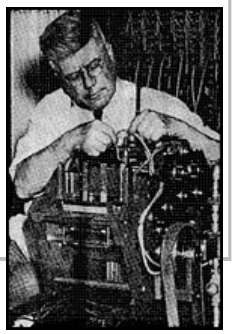
The mainframe tape subsystem was upgraded to introduce Virtual Tape Technology. This technology allows all mainframe tape data to be placed on modern high capacity tape under robotic control thus eliminating the requirement for provide storage racks for the agency's 75,000 mainframe tapes. This upgrade also automatically provides a second redundant copy of all tape data to make tape I/O errors a thing of the past.

The following chart illustrates the increasing use of computers at Statistics Canada, mirroring their growing use in the large business community.



A look at our past

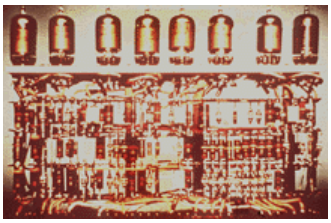
Resident inventor designs Bureau's mechanical tabulators



Fernand Bélisle, born in the Eastern Townships of Quebec in 1889, became a legend in the DBS [Dominion Bureau of Statistics] as a mechanical genius. The 1941 administrative report of the Dominion Statistician [R.H. Coats] cited Bélisle as the developer of the Pantograph machine, used since 1911 for punching cards; and as the inventor of the electric gang-punch machine used so effectively in the 1931 Census, and of the compressed-air sorter-counter machine that "tabulated a record amount of information in the 1941 Census." Referring to the compressed-air tabulator, Coats told the 1940 Conference on Canadian-American Affairs that the DBS "has developed a census machine which is the envy of the world." Apart from a couple of early breaks in service, Bélisle worked for the Bureau from 1911 until he retired in 1950. He died in 1963.

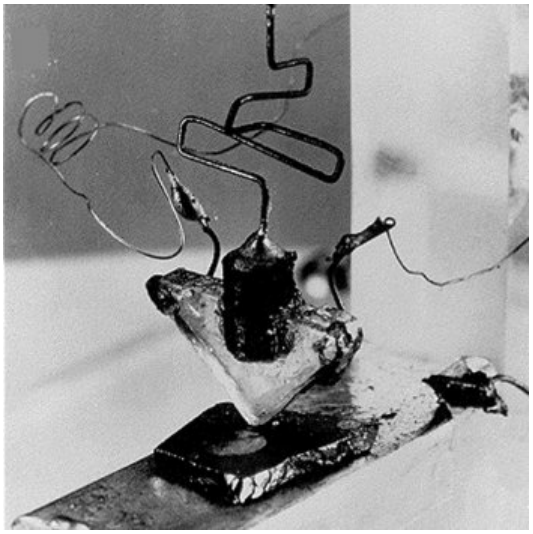
A general view of the *IBM 705 III* computer in 1969

Vacuum tubes



Microprocessor

The first transistor



Date Modified: 2013-07-23

The computer industry

Commercial installations

Large companies and businesses don't usually have just one large computer. Typically, they have a range of specialized computers that are networked. Depending on the size of the company, there could be one or two very powerful mainframe computers, a number of small, medium and large Unix-based midrange computers, and many individual personal computers (PCs).

In fact, most organizations' computers are networked. The network that connects computers in one location is usually referred to as a local area network (LAN); connected individual LANs are a wide area network (WAN). A WAN is made up of servers, workstations, a network operating system, and a communications link. Servers are high-speed machines that hold programs and data shared by network users. Using this network, authorized personnel can access data and make use of computer facilities anywhere in the distributed enterprise.

Until 15 or 20 years ago, mainframes were the most common computers in business. Today's mainframe computers generally run older computer applications (often called 'legacy' systems) that were written several years ago. Depending on the system requirements, these applications are designed to operate on the mainframe, midrange computers (servers) or networked personal computers.

Currently, with appropriate authorization, users can access data files via their computer, regardless which type of operating system their data resides on.

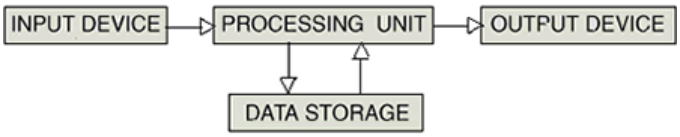
The Internet and intranet

The Internet has revolutionized the computer and communications world like nothing before it. The invention of the telegraph, telephone, radio, and computer set the stage for this unprecedented integration of capabilities. The Internet is a medium for worldwide broadcasting, and for collaboration and interaction between individuals and their computers without regard to geographic location or time restraints.

Today businesses, academic institutions and people at home use the Internet to send e-mail messages around the world, and to conduct research. However, most businesses consider the Internet too slow, unreliable and insecure for widespread internal use. To solve this problem, they have set up their own private internal internets, known as intranets. These intranets provide the convenience of the Internet, along with the performance and security of an in-house system. Intranets can be connected to the public Internet via secure gateways, which have 'firewalls' to prevent unwanted external access to internal systems.

Computer hardware

To function properly, a computer system needs the following hardware components:



- An input device allows users to enter data or program the computer.
- The processing unit controls all activities within the system.
- Data Storage holds databases, files and programs.
- Output devices present the finished information product to the user.

Input device

Data can be entered into a computer through a range of ways, including:

- scanner, keyboard, mouse;
- floppy disk or CD-ROM;
- magnetic tape, light pen or barcode reader, microphone and digital camera;
- communication and telephone line; and
- Internet e-mail and database.

Processing unit

The Central Processing Unit (CPU) is the heart of a computer system. In most modern computers, the CPU consists of just one or two silicon chips that are small enough to hold in one hand, but contain many millions of logic circuits. A CPU can execute millions of instructions per second.

Associated with the CPU is the random-access memory (RAM). The RAM has to be big enough to hold all programs and data that are being worked on at any given time. RAM size ranges from a few million bytes to a few hundred million bytes.

Output devices

Disk drives are used by computers to store data. A current PC might have a 2.0-gigabyte (GB) disk drive (20 billion bytes), while large computers might have a number of disk drives holding tens or hundreds of gigabytes.

Data can be copied onto magnetic tape or CD-ROMs for backup, transfer between computers, or long-term storage. A CD-ROM can hold 700 megabytes (MB), while tape units can range from about 500 MB to 24 GB. Data can also be copied onto diskettes (floppy disks), but their small size (1.44 MB) limits their usefulness with large data sets.

Output devices

An output device is any peripheral device that presents output from the computer. Output devices include:

- video monitors;
- various sorts of printers;
- magnetic disks and tapes;
- CD-ROMs;
- data communication and telephone lines; and
- stereo speakers.

Computers can also be used to drive industrial processes, control chemical plants, and lock/unlock security doors. Modern car-engine management systems, office elevators, VCRs, and numerous other domestic and industrial systems are now controlled by miniature computer systems.

Storage and retrieval

Much of the computer's power comes from its ability to store, sort and classify data. Over the past few years, disk systems have become very inexpensive and reliable and it is now possible to obtain disk systems that will store billions of bytes of data for just a few hundred dollars. This thousand fold drop in price has greatly improved the usefulness of computers. It is now possible to hold all of a company's information, going back for years, on a computer.

Software

There are two basic types of computer software: systems software, which controls the operation of the computer, and applications software, which performs useful tasks for the user.

Systems software itself is divided into two classes: operating systems, and tools and utilities.

An **operating system** generally has two levels:

- The *lower-level basic input-output system* (BIOS) controls the most basic functions, such as: reading and writing RAM (memory), and input to and output from peripherals such as mouse, keyboard, printer and screen.
- The *higher-level main operating system* (e.g., Windows) acts as a platform to host programs. It provides the user interface to control the computer's operation, and the environment to effectively operate application software. For example, it provides a file subsystem with its structure of drive names, directories, folders, files, and indexes; and file-handling facilities such as creating, copying and deleting.

Typical operating systems include DOS, UNIX, Mac OS and Windows. Mac OS and Windows have a user-friendly graphical user interface (GUI) which enables computer control by means of windows, menus, icons and a mouse. DOS and UNIX require the user to type precise commands, which can be hard to remember. Windows, Mac OS and UNIX can run many programs at one time (multiprogramming), which makes for more efficient use of computer hardware and more convenience for the user.

Tools and utilities software are usually necessary to make productive use of a computer. Some of the software are provided with the operating system, while others can be downloaded or purchased separately. Typical system utilities include Internet browsers, antivirus software, program compilers, editors and file-backup systems.

Applications software can also be divided into two classes: personal productivity tools and other computer applications.

Personal productivity tools

are commercial products designed to handle standard computing tasks such as word processing, numerical analysis, data manipulation and storage, and data presentation. Typical products include:

- **Word processing**
  Word processing software is designed to create documents such as letters, reports, newspaper articles and manuscripts. It was one of the first applications available for personal computers, which helped streamline large amounts of routine typewriting. This type of software succeeded because it allowed text to be edited without having to retype the whole document. MS Word, WordPerfect, and Word Pro are some examples of word processing products.
- **Spreadsheets**
  Spreadsheets, used by bookkeepers and office managers to organize business information and perform electronic accounting, are very useful for handling tabular data. Addition, subtraction, division, multiplication and totaling can be done very quickly, and all results can be automatically recalculated later if new data are inserted. Formatting and graphing facilities are used to aid analysis and presentation. Hundreds of functions enable typical statistical, engineering, economics and business calculations to be performed automatically (e.g., compound interest, standard deviation). Well-known spreadsheet packages are MS Excel and Lotus 123.
- **Databases**
  Database packages make it easy to organize and store data in a uniform fashion. Data can be quickly and systematically searched, sorted and presented. Databases can be used by people with no special training to create mailing lists or record store inventories, and they can also be used by professional programmers to produce complex applications to help run a business. MS Access and Lotus Approach are database packages.
- **Presentation**
  Presentation packages are used to illustrate discussions and lectures, replacing hand-drawn or typed overhead projector slides. In many presentations

and lectures given today, the presenter plugs a laptop computer directly into a projector to show slides on a screen. MS PowerPoint and Lotus Freelance are examples of presentation packages.

- **Graphics**
  Graphics packages enable users to create drawings, paint pictures and enhance or manipulate scanned images such as photographs or artwork. MS Paint and Corel Draw! are examples of graphics packages.

- **Desktop publishing**
  Desktop publishing packages are intended to enhance the final appearance or layout of text and graphics, to make them suitable for publishing. Graphics can come from a library of clip art, or be created by graphics packages. Some typical desktop publishing packages are MS Publisher, Paint Shop Pro, and PageMaker.

Other computer applications

Personal productivity tools are used at home, at school and in the office. These tools are fairly inexpensive and available on many computers. However, businesses and organizations usually buy computers to automate major business functions, not handled by personal productivity software.

Some software applications can cost many thousands of dollars to buy or develop, while a major banking or airline reservation system could cost millions. Application software can be purchased 'off the shelf' or developed for a specific purpose. An accounting package is an example of a purchased application, while a system to handle parking fines might be designed and written from scratch.

An example of a complex computer application is an airplane's autopilot navigation system. This system's software receives information (e.g., compass heading and Global Positioning System data) and outputs data that controls rudder and flap hydraulics which adjust the course of the plane.

Systems analysts, programmers and users

If it is decided to develop software to automate a task, the work is done by systems analysts and programmers.

Systems analysts

Systems analysis is the process of breaking down a data-processing problem into functional components to determine the best method of handling the problem. Systems analysts design and modify systems by turning user requirements into a set of functional specifications, the blueprint of the system.

In consultation with the user, the systems analyst must:

- define the system problems of an organization;
- investigate the current situation to determine new system requirements;
- develop the specifications that are practical, efficient, cost-effective and make the best use of available hardware and software;
- communicate the new system requirements to all parties concerned; and
- assist in implementing the new system.

Programmers

Programming is the process of producing a set of instructions to make a computer perform a specified activity. Programmers take system analysis results and develop computer programs to solve the problem.

In consultation with the user, a programmer must:

- understand problems and plan solutions;
- design programs using data flow diagrams and other design tools;
- write programs to implement the design;
- test programs and correct any problems; and
- write detailed documentation of programs and their operation.

Users

The user is the final judge of whether a computer system is meeting the needs it was designed to fulfill. The better the link between the automated system components and the user, the more likely it is that the system will be effective. Modern system designers consult widely with users in order to design systems that meet their needs, and designers put considerable effort and ingenuity into designing the interface between systems and users.

Exercises

1. Computer systems are a combination of three ingredients. What are they?

2. Name three hardware components of a computer system.

3. Give two reasons why large organizations set up their own internal intranet system.

4. Why would someone doing data analysis find spreadsheets useful?

5. Why is it important that a systems analyst consult with the system user before trying to design a new system?

6. Where was the first electronic computer developed?

7. What was the first calculator called?

8. What is the difference between "Intranet" and the Internet?

Answers

1. Computer systems are a combination of hardware, software and users.

2. The three hardware components include:
    a. Input devices: diskettes, CD-ROMs, magnetic tape, light pens, barcode readers, microphones, digital cameras, telephone lines, other communication lines (cable, etc.).
    b. Data storage devices: diskettes, CD-ROMs, magnetic tape, disk drives.
    c. Output devices: video monitors, printers, magnetic tape, diskettes, CD-ROMs, data communication lines, telephone lines, stereo speakers.

3. The two reasons why large organizations set up their own intranet are to get the convenience of the Internet with the performance and security of an in-house system.

4. A data analyst would find spreadsheets very useful because they automatically handle tabular data.

5. It is important that a systems analyst consult a systems user:
    a. to identify problems in the organization's system.
    b. to determine new system requirements.
    c. to design a new system that is practical, efficient, and cost effective, and that makes the best use of the available hardware and software.

6. In Britain, by Thomas Flowers.

7. An Abacus.

8. An internal 'Intranet' network is a microcosm of the Internet set up within a large enterprise to facilitate work. Intranets are more secure and reliable than the Internet and they have "firewalls" to prevent unwanted external access.

Bibliography

The following is an alphabetical listing of all research and resource material used in the development of *Statistics: Power from Data!*

A

Agriculture and Agri-food Canada, Research Branch. *Types of Data*. March 23, 2001. (Accessed January 22, 2002).

Aloha-Sidaway, Janice and Margaret McKinnon. 1999. "How to assess the pedagogical soundness of new technology." *So You Want to Develop an Effective Learning Resource: A collection of ideas*. Hull: Minister of Canadian Heritage.

Anderson, Jon E. "Chapter 3: Producing Data: Systematic Sampling." *Course Syllabus for Statistics 1601*. University of Minnesota. February 17, 1999. (Accessed May 4, 2001).

Australian Bureau of Statistics. 1990. *Australia: Working It Out! Core material for Australian studies*. Queanbeyan, AU: Australian Bureau of Statistics.

B

Black, Paul E. "The Role of Graphics." *Engineering Statistics Handbook*. Beta version. National Institute of Standards and Technology (NIST). September 4, 1998. (August 7, 2001).

Bock, Rudolf K. and W. Krischer. "Stratified Sampling." *The Data Analysis Briefbook*. Version 16. April 7, 1998. (Accessed May 4, 2001).

Brady, Bennett M. and Joseph W. Duncan. 1978. *Statistical Services in Ten Years' Time*. Oxford, UK: Pergamon Press.

Broster, Eric James. 1974. *Glossary of Applied Management and Financial Statistics*. Essex, UK: Gower Press.

C

Cailliau, Robert. *A Little History of the World Wide Web*. October 3, 1995. (Accessed November 26, 1999).

Casey, Nancy and Mike Fellows. "Graphs, Games, and the NCTM Standards." *MegaMathematics*. Los Alamos National Laboratory, Computer Research and Applications Group. 1998. (Accessed May 4, 2001).

Center for Support of Teaching and Learning. "Constructing Bar Graphs." *Self-instructional Mathematics Tutorials*. Syracuse University. 2000. (Accessed September 6, 2001).

Clarion Research. *Quality Control Procedures*. 1998. (Accessed September 13, 2001).

Crowson, Scott. "Hot Economy Stalls Murder Rate: Calgary Second Lowest in Country." *The Calgary Herald*. Thursday, October 19, 2000. (News, B1).

D

Data Interpretation Workshop. 1998. "Effective Publication Graphics, Course 04482." *Course Calendar for September 8 to October 20, 1998*. Ottawa: Minister Responsible for Statistics Canada. (Statistics Canada Catalogue no. 10H0053GPE).

Deller, Stephen. 1988. "The History of the Canadian Census." *Families*. Vol. 23, No. 3. Toronto: Ontario Genealogical Society.

Doucet, Ed and Edvard Outrata. 1998. "Impact of New Processing Techniques on the Management and Organization of Statistical Data Processing." *Journal of the United Nations Economic Commission for Europe*. ECE 5: 201–221.

Drott, Carl M. *Dr. Drott's Random Sampler: Using the Computer as a Tool for Library Management*. College of Information Science and Technology, Drexel University. 1994. (Accessed June 11, 2001).

E

Easton, Valerie J. and John H. McColl. "Presenting Data." *STEPS Statistics Glossary Web*. Version 1.1. September 1997. (Accessed May 4, 2001).

Erickson, Bonnie H. and T.A. Nosanchuk. 1992. *Understanding Data*. Second Edition. Toronto: University of Toronto Press.

F

Fellegi, Ivan P. and D. Holt. 1976. *A Systematic Approach to Automatic Edits and Imputation*. Ottawa: Minister Responsible for Statistics Canada. (Statistics Canada Catalogue no. STC3432B).

Fellegi, Ivan P. 1967. *Computer Methods for Geographical Coding and Retrieval of Data in the Dominion Bureau of Statistics*. Ottawa: Dominion Bureau of Statistics, Census Division. (Statistics Canada Catalogue no. STC3038).

Fellegi, P. Ivan and Simon A. Goldberg. 1971. *The Computer and Government Statistics*. Ottawa: Minister Responsible for Statistics Canada. (Statistics Canada Catalogue no. STC2545).

Ferber, Robert. 1980. *What is a Survey?* Subcommittee of the Section on Survey Research Methods, American Statistical Association (ASA). Washington, DC: ASA.

Filliben, James. "Scatterplot." *Engineering Statistics Handbook*. Beta version. National Institute of Standards and Technology (NIST). 1998. (Accessed August 7, 2001).

Financial Forecast Centre. *What is the Correlation Coefficient?* 10 Year T-Note Forecast. Applied Reasoning Incorporated. 1997. (Accessed November 25, 2002).

Friedman, Herbert. 1972, Introduction to Statistics. New York: Random House.

G

Galloway, Allison. *Sampling: A Workbook*. August 25, 1997. (Accessed June 11, 2001).

Garson, David G. "Sampling." Syllabus for PA 765: *Quantitative Research in Public Administration*. North Carolina State University. Fall 2001. (Accessed May 31, 2002).

Gower, Allen. 1997. *Questionnaire Design: Statistics Canada, Course 410E*. Statistics Canada, Questionnaire Design Resource Centre. Ottawa: Minister Responsible for Statistics Canada. (Statistics Canada Catalogue no. 10H0046XPE).

Gower, Allen and Karen Kelly. 1993. *How Big Should the Sample Be?* Ottawa: Minister Responsible for Statistics Canada. (unpublished).

Gravetter, Frederick J. and Larry B. Wallnau. 1999. *Essentials of Statistics for the Behavioral Sciences*. Third Edition. Pacific Grove, CA: Brooks/Cole Publishing Company.

H

H & H Servicco Corporation. "Glossary of Terms." *Sampling Plans*. January 15, 2002. (Accessed February 21, 2002).

Han, Kyunghee. "Selection of a Sample." *EDRS 1: Educational Research I*. University of Mississippi. Spring 2001. (Accessed June 11, 2001).

Haralambos, Micheal and Martin Holborn. 1987. *Sociology: Themes and Perspectives*. London, UK: Collins Educational. (Pages 734–740 excerpted on website) 2000. (Accessed February 28, 2002).

Hauser, Philip M. *Social Statistics in Use*. 1975. New York: Russell Sage Foundation.

Huff, Darrell. 1954. *How to Lie with Statistics*. New York: W. W. Norton and Company.

Human Resources Development Canada. Labour Force Survey – *Definitions and Explanations*. 2002. (Accessed January 25, 2002).

J

Johnson, Richard Arnold and Jane F. Gentleman. 1989. *Mainframe SAS Enhancements in Support of Exploratory Data Analysis*. Statistics Canada, Analytical Studies Branch. Ottawa: Minister Responsible for Statistics Canada. (Statistics Canada Catalogue no. 11F0019E no. 24).

Jones, James. "Estimating the Population Mean." *Lecture Notes for Math 170: Introduction to Statistics*. Richland Community College. 2001. (Accessed May 4, 2001).

— "Introduction to Estimation." *Lecture Notes for Math 170: Introduction to Statistics*. 2001. (Accessed May 4, 2001).

— "Introduction to Probability." *Lecture Notes for Math 170: Introduction to Statistics*. 2001. (Accessed May 4, 2001).

— "Measures of Central Tendency." *Lecture Notes for Math 170: Introduction to Statistics*. 2001. (Accessed May 4, 2001).

— "Stats: Estimating the Mean." *Lecture Notes for Math 170: Introduction to Statistics*. 2001. (Accessed May 4, 2001).

— "Stats: Frequency Distribution and Graphs." *Lecture Notes for Math 170: Introduction to Statistics*. 2002. (Accessed May 4, 2001).

K

Kaushal, Ritu and Normand J. D. Laniel. 1993. *Computer Assisted Interviewing Data Quality Test*. Ottawa: Minister Responsible for Statistics Canada. (Statistics Canada Catalogue no. 11N022).

Kendall, Sir Maurice G. and William R. Buckland. 1982. *A Dictionary of Statistical Terms*. Fourth Edition, revised and enlarged. London, UK: Longman Group.

Key, James P. "Module R8–Sampling." *Research Design in Occupational Education*. 1997. (Accessed September 13, 2001).

Kirkman, Tom. "Descriptive Statistics." *Tools for Science*. 2001. (Accessed June 11, 2001).

Klinke, Sigbert. 1997. *Data Structures for Computational Statistics*. Berlin: Physica-Verlag.

Kristula, Dave. *A History of the Internet*. August 2001. (September 9, 2001).

Krotki, Karol J. 1977. *Census of Computer Software at Statistics Canada*. Ottawa: Minister Responsible for Statistics Canada. (Statistics Canada Catalogue no. STC1210).

L

Lane, David M. "Arithmetic Mean." *Hyperstat Online*. 1993. (Accessed November 24, 2000).

— "Box Plot." *Hyperstat Online*. 1993. (Accessed November 24, 2000).

— "Linear Relationship." *Hyperstat Online*. 1993. (Accessed November 24, 2000).

— "Nominal Scale." *Hyperstat Online*. 1993. (Accessed November 24, 2000).

— "Sampling Fluctuation." *Hyperstat Online*. 1993. (Accessed September 13, 2001).

— "Scatterplots." *Hyperstat Online*. 1993. (Accessed November 24, 2000).

— "Stem and Leaf Plots." *Hyperstat Online*. 1993. (Accessed November 24, 2000).

Leiner, Barry M. et al. *A Brief History of the Internet*. Version 3.31. Internet Society (ISOC). August 4, 2000. (September 9, 2001).

Lemeshow, Stanley and Paul S. Levy. 1991. *Sampling of Populations: Methods and Applications*. Second Edition. New York: John Wiley & Sons.

Lemeshow, Stanley and Paul S. Levy. 1999. *Sampling of Populations: Methods and Applications*. Third Edition. New York: John Wiley & Sons.

Lohr, Sharon L. 1999. "Chapter 6: Sampling with Unequal Probabilities." *Sampling: Design and Analysis*. Pacific Grove, CA: Duxbury Press.

Loveday, Robert. 1958. *A First Course in Statistics*. Cambridge, UK: Cambridge University Press.

M

Martin, Marjorie. 1986. *Introduction to Interviewing*. Statistics Canada, Survey Operations Division. Ottawa: Ministry Responsible for Statistics Canada. (Statistics Canada Catalogue no. STC1595).

Mathieu, Richard G. and Omar Khalil. "Data Quality in the Database Systems Course." *Data Quality*. 2001. (Accessed September 13, 2001).

Menzies, Heather. 1982. *Computers on the Job: Surviving Canada's Microcomputer Revolution*. Toronto: James Lorimer and Company.

Meyers, Jeremy, *A Short History of the Computer*. November 3, 1999. (Accessed November 26, 1999).

Murdock, Everett. *History, the History of Computers, and the History of Computers in Education*. California State University Long Beach. 1994. (Accessed November 2, 1999).

N

National Statistics Office, Republic of the Phillipines. *Technical Notes on the Seasonal Adjustment of CPI*. June 25, 2001. (Accessed January 25, 2002).

Neiswanger, William Addison. 1947. *Elementary Statistical Methods: As Applied to Business and Economic Data*. Revised Edition. New York: The Macmillan Company.

Niles, Robert. "Margin of Error." *Statistics Every Writer Should Know*. 1996. (Accessed September 13, 2001).

Noether, Gottfried. 1991. *Introduction to Statistics: A Fresh Approach*. New York: Springer-Verlag.

NVision Research. *Analysis Capabilities*. 2001. (Accessed September 13, 2001).

O

Official site of the Canadian Football League. (Accessed September 26, 2002).

Ottawa Citizen. "Crime by the Numbers." *Ottawa Citizen*. Tuesday, July 25, 2000. (Editorial, A14).

P

Parlor, Margaret. 1991. *Canada's International Trade Statistics: Yesterday, Today and Tomorrow*. Ottawa: Minister Responsible for Statistics Canada. (Statistics Canada Catalogue no. 65-001).

Paton, David G. 1996. "Editing Strategies and Systems Used by the Canadian General Social Survey: Their Evolution over Ten Years of Data Collection and Processing." *Survey and Statistical Computing 1996*. Ottawa: Minister Responsible for Statistics Canada. (Statistics Canada Catalogue no. 12F0066XPE).

Plonsky, M. "Frequency Distributions." *Psychological Statistics*. 1997. (Accessed January 22, 2002).

Polsson, Ken. *Chronology of Personal Computers*. 1994. (Accessed November 26, 1999).

R

Renckly, Tom. "Chapter 4: Sampling Techniques and Related Statistical Concepts." *Sampling and Surveying Handbook*. Air University. June 7, 2001. (Accessed June 11, 2001).

Routio, Pentti. "Sampling." *Arteology*. May 10, 2001. (Accessed June 11, 2001).

S

Sanders, Donald A. et al. 2001. *Statistics: A First Course*. First Canadian Edition. Whitby, Ontario: McGraw-Hill Ryerson.

Satin, Alvin and Wilma Shastry. March 1983. *Survey Sampling: A Non-mathematical Guide*. Second Edition. Statistics Canada, Social Survey Division. Ottawa: Minister Responsible for Statistics Canada. (Statistics Canada Catalogue no. 12-602E)

Sharov, Alexei. "2.4: Stratified Sampling." *Lecture Notes for Quantitative Population Ecology*. Virginia Tech University, Department of Entomology. December 1, 1996. (Accessed June 11, 2001).

Sheskin, Ira M. 1985. *Survey Research for Geographers*. Washington: Association of American Geographers.

Shodor Education Foundation. *What? Pie Chart*. 1997. (Accessed May 4, 2001)

Stark, Philip B. "Chapter 1: Statistics." *Statistics Tools For Internet and Classroom Instruction with a Graphical User Interface (SticiGui)*. University of California, Berkeley. 2001. (Accessed May 4, 2001).

— "Chapter 2: Measures of Location and Spread." *SticiGui*. University of California, Berkeley. 2001. (Accessed May 4, 2001).

— "Chapter 2: Measures of Location." *SticiGui*. University of California, Berkeley. 2001. (Accessed May 4, 2001).

Statistics Canada and the Social Science Federation of Canada. 1983. *Historical Statistics of Canada*. Second Edition. Ottawa: Minister Responsible for Statistics Canada. (Statistics Canada Catalogue no. 11-516E).

Statistics Canada. "Canada, Labour Force Characteristics, Monthly from January 1976, Unadjusted." *CANSIM*. 1999. (Accessed August 23, 1999).

Statistics Canada. 1970. *Content of Questionnaire for the 1971 Census of Canada*. Statistics Canada, Census Division. Ottawa: Minister Responsible for Statistics Canada. (Statistics Canada Catalogue no. STC1623E).

Statistics Canada. 1980. *Finding and Using Statistics: A Basic Guide from Statistics Canada*. Ottawa: Minister Responsible for Statistics Canada. (Statistics Canada Catalogue no. STC2530E).

Statistics Canada. 1980. *Statistics Canada Catalogue*. Statistics Canada, User Advisory Services Division. Ottawa: Minister Responsible for Statistics Canada. (Statistics Canada Catalogue no. 11-204E).

Statistics Canada. 1987. *Statistics Canada Quality Guidelines*. Second Edition. Ottawa: Minister Responsible for Statistics Canada. (Statistics Canada Catalogue no. 12-539-XPE).

Statistics Canada. 1987. *Statistics Canada's Policy on Informing Users of Data Quality and Methodology*. Ottawa: Minister Responsible for Statistics Canada. (Statistics Canada Catalogue no. STC3171).

Statistics Canada. 1991. *Focus for the Future*. Ottawa: Minister Responsible for Statistics Canada. (Statistics Canada Catalogue no. STC2537E).

Statistics Canada. 1991. *Historical Labour Force Statistics*. Ottawa: Minister Responsible for Statistics Canada. (Statistics Canada Catalogue no. 71-201-XPB 2000).

Statistics Canada. 1992. 1991 *Census Handbook*. Ottawa: Minister Responsible for Statistics Canada. (Statistics Canada Catalogue no. 92-305E).

Statistics Canada. 1993. 75 Years and Counting: *A History of Statistics Canada*. Ottawa: Minister Responsible for Statistics Canada. (Statistics Canada Catalogue no. 11-531).

Statistics Canada. 1994. *Revised Intercensal Population and Family Estimates, July 1, 1971–1991*. Ottawa: Minister Responsible for Statistics Canada. (Statistics Canada Catalogue no. 91-537).

Statistics Canada. 1996. *Questionnaire content: 1996 Census of Population*. Ottawa: Minister Responsible for Statistics Canada. (Statistics Canada Catalogue no. 92N0064XPE).

Statistics Canada. 1997. *1996 Census Handbook*. Ottawa: Minister Responsible for Statistics Canada. (Statistics Canada Catalogue no. 92-352-XPE).

Statistics Canada. 1997. *Guide to the Labour Force Survey*. Ottawa: Minister Responsible for Statistics Canada. (Statistics Canada Catalogue no. 71-543-GIE).

Statistics Canada. 1998. *Case Studies for Business Schools: Preliminary Review of Existing Statistics Canada 'Case-Like' Materials*. Statistics Canada, Marketing Division. Ottawa: Minister Responsible for Statistics Canada.

Statistics Canada. December 2, 1999. *Statistics Canada Mainframe Computers, Since 1960*. Revised Edition. Ottawa: Minister Responsible for Statistics Canada.

Statistics Canada. 1998. *Statistics Canada Quality Guidelines*. Third Edition. Ottawa: Minister Responsible for Statistics Canada. (Statistics Canada Catalogue no. 12-539-XIE 1998).

Statistics Canada. 1998. *Surveys from Start to Finish (Workshop)*. Ottawa: Minister Responsible for Statistics Canada. (unpubished).

Statistics Canada. 1998. *The How and Why of Business Statistics: the Respondent's Perspective*. Respondent Relations Unit, Communications Division. Ottawa: Minister Responsible for Statistics Canada. (Statistics Canada Catalogue no. 10F0213XPE).

Statistics Canada. 1999. *Canada Year Book 1999*. Ottawa: Minister Responsible for Statistics Canada. (Statistics Canada Catalogue no. 11-402-XPE).

Statistics Canada. 2000. *Education Indicators in Canada: Report of the Pan-Canadian Education Indicators Program, 1999*. Ottawa: Minister Responsible for Statistics Canada. (Statistics Canada Catalogue no. 81-582-XPE 1999).

Statistics Canada. 2000. *Human Activity and the Environment 2000*. Ottawa: Minister Responsible for Statistics Canada. (Statistics Canada Catalogue no. 11-509-XPE).

Statistics Canada. 2000. Survey Skills Development Course, SSDC-64. Ottawa: Minister Responsible for Statistics Canada. (Statistics Canada Catalogue no. 12N0079XPE).

Statistics Canada. *Canada Year Book 1999*. (Teachers' Kit) February 18, 2002. (Accessed May 29, 2002).

Statistics Canada. *Graphing in the information age*. (Teachers' Kit). February 12, 2002. (Accessed March 23, 2000).

Statistics Canada. *Labour Force Survey Questionnaire*. Version 1997. 1997. (Accessed October 28, 1999).

Statistics New Zealand: Te Tari Tatau. "Good Stem and Leaf Graphs." *School's Corner*. 1998. (Accessed November 23, 1998).

Strata.com. *Statistical Graphs and Display*. 2001. (Accessed February 7, 2001).

Suess, Jack. *Classifying Variables by Type*. University of Maryland. January 15, 1996. (Accessed May 4, 2001).

T

Texas Instruments. 1980. *Student Calculator Math: An Expanded Edition of TI's Popular "Great International Math on Keys" Book*. Texas: Texas Instruments Incorporated Learning Center.

Twelker, Paul A. *Lesson 15: Discrete Random Variables*. Trinity International University. March 21, 2000. (Accessed May 4, 2001).

V

Vanier Institute of the Family. 1994. *Profiling Canada's Families*. Ottawa: Vanier Institute of the Family.

W

Wattier, Mark, J. "Descriptive Statistics." *POL 660 Course Syllabus*. Murray State University. 2001. (Accessed November 24, 2000).

Webopedia.com. "Column Graph." *Webopedia*. INT Media Group. September 1, 1996. (Accessed May 4, 2001).

— "Raw Data." *Webopedia*. INT Media Group. September 1, 1996. (Accessed January 22, 2002).

Westgard, James O. *The Idea of Statistical Quality Control*. 2000. (Accessed September 13, 2001).

Weston, Harley. "A note on standard deviation." *Math Central*. July 14, 1997. (Accessed June 11, 2001).

White, Stephen. *A Brief History of Computing*. November 8, 1999. (Accessed November 26, 1999).

Wilhelm, E., R. Dibbs and Wilma Shastry. 1983. *Definitions of Terms Used in Survey Research*. Ottawa: Minister Responsible for Industry Canada.

Wilson, Terry C. 1980. *Researcher's Guide to Statistics: Glossary and Decision Map*. Maryland: University Press of America.

Worton, David Albert. 1998. *The Dominion Bureau of Statistics: A History of Canada's Central Statistics Office and its Antecedents, 1841-1972*. Kingston: McGill-Queen's University Press. (Statistics Canada Catalogue no. 12-582-XPE).

Wright, Jane-Marie. *MA 22 Statistics Unit: Creating a Frequency Table and Histogram*. Suffolk County Community College, Mathematics Department. February 25, 1999. (Accessed February 19, 2002).

Y

Youtsey, Carolyn. "Line Graphs." *Carolyn's Unit on Graphing*. University of Illinois at Urbana-Champaign. August 27, 1998. (Accessed May 4, 2001).

Statistics: Power from Data!
Glossary

The definitions below provide information for those who have questions about statistics but who do not need highly technical explanations.

The definitions provided here are, in some cases, oversimplifications of highly complex concepts. For those interested in more technical definitions, click here for a list of Statistics Canada's dictionaries and definitions.

A

Aboriginal peoples

Persons that are North American Indian, Métis or Inuit (Eskimo).

Accessibility

Accessibility reflects the availability of information from the holdings of the agency. It takes into account the suitability of the format in which the information is available; the media of dissemination; the availability of metadata (descriptive text); and whether the user has reasonable opportunity to know it is available and how to access it. For users, the affordability of the information in relation to its value to them is also an aspect of this characteristic.

Accuracy

The extent to which the results of a calculation or reading of an instrument approach the true values of the calculated or measured quantities and are free from error. (See also: precision.)

Administrative by-product

Data available from the information recorded in administrative records, applications, reports, etc.

Age–sex pyramid

A graph designed to represent the age structure of a population. It consists of two horizontal histogram graphs joined together. Example of an age-sex population pyramid

B

Baby boomer

Generally, those persons born following World War II between the years 1946 and 1966.

Balance of payments

The balance of payments covers all economic transactions between Canadian residents and non-residents. It includes the current account and the capital and financial account. **Source:** ARCHIVED - *The Daily*-Friday, May 28, 2004

Bar graph

A diagram that compares bars of the same width but of different heights according to the statistics or data they represent. Bar graphs are horizontal. (A vertical bar graph is called a column graph.) Example of a bar graph

Batch keying

One of the oldest methods of data capture. It uses a computer keyboard to type in the data. This process is very practical for high-volume entry where fast production is a requirement. No editing procedures are necessary but there must be a high degree of confidence in the editing program. Also, validity and range edits need to be implemented to ensure quality keying. This does not mean the data are being re-edited, but if a field is numeric and alpha characters are entered instead, the error will be flagged. This approach can be beneficial when used for large surveys with many questions and edits.

Bias

In estimation, the bias refers to the value of a parameter of a probability distribution, the difference between the expected value of the estimator and the true value of the parameter.

C

Capital and financial account

The capital and financial account mainly comprises of transactions in financial instruments. Financial assets and liabilities with non-residents are presented under three functional classes: direct investment, portfolio investment and other investment. These investments belong either to Canadian residents (Canadian assets) or to foreign residents (Canadian liabilities). Transactions resulting in a capital inflow are presented as positive values, while capital outflows from Canada are shown as negative values. **Source:** ARCHIVED - *The Daily*-Friday, May 28, 2004

Categorical data

Consists of data that can be grouped by specific categories (also known as qualitative variables). Categorical variables may have categories that are naturally ordered (ordinal variables) or have no natural order (nominal variables). For example, the variable "height" is ordinal because it contains the categories "short", "average" and "tall" which are naturally ordered according to ascending height. On the other hand, variables such as "sex" and "hair colour", which have no natural category order, are examples of nominal variables.

Census

The collection of information about all units in a population, sometimes also called a 100% sample survey. (When capitalized, "Census" usually refers to the national Census of Population.)

Central processing unit

The Central Processing Unit (CPU) is the heart of a computer system. It can be small enough to hold in your hand but can contain millions of logic circuits.

Central tendency

A measure of location of the middle or the centre of a distribution. Central tendency can refer to a wide variety of measures such as mean, median and mode. The mean is the most commonly used measure of central tendency.

Characteristic

A property which helps to differentiate between items of a given population. This differentiation may be either qualitative or quantitative.

Class intervals

If a variable has a large number of values, it is easier to present the data by grouping the values into class intervals (i.e., age of the population presented as age groups, for example 0 to 4, 5 to 9, 10 to 14, 15 to 19, etc.) rather than presenting all of the values together. This makes it easier to see the trends in the data.

Classification

The organized representation of a given population into homogeneous categories.

Cluster

A set of units grouped together on the basis of some well-defined criteria. The cluster may be an existing grouping of the population such as a city block, or hospital; or conceptual such as the area covered by a grid imposed on a map.

Coding

A process for converting questionnaire information into numbers or symbols to facilitate subsequent data processing operations. Sometimes, this involves interpreting responses and classifying them into predetermined results.

Coefficient of determination

A measure of how much the variability of one given variable depends on its relationship with another given variable. It is calculated by squaring the value of the linear correlation, **r**. For example, in a linear model, a correlation of 0.80 (**r** = 0.80) would mean that $\mathbf{r^2}$, the coefficient of determination, would equal 0.64. Therefore, 64% of the variability in the Y values could be predicted based on the relationship with the X values.

Coefficient of variation

A measure of dispersion calculated by dividing the standard deviation of a distribution divided by its mean. The standard error of an estimate, expressed as a ratio or percentage of the estimate.

Coherence

Coherence reflects the degree to which the data and information from a single statistical program are brought together with other data information and are logically connected and completed. Fully coherent data are consistent—internally, over time, and across products and programs. Where applicable, the concepts and target populations used or presented are logically distinguished from similar concepts and target populations or from commonly used notions or terminology.

Cold deck

Makes use of a fixed set of values, which covers all of the data items. These values can be constructed with the use of historical data, subject-matter expertise, etc. A 'perfect' questionnaire is created in order to answer complete or partial imputation requirement.

Column graph

A vertical bar graph. Example of a column graph

Common-law

Two people of the opposite or of the same sex who live together as a couple, but who are not legally married to each other.

Confidence interval

An estimate using a range of values (an interval) to predict the expected value of an unknown parameter, accompanied by a specific level of confidence, or probability, that the estimate will be correct (i.e. that the interval will in fact contain the true value of the parameter).

Confidentiality

In a confidential survey, the privacy of information provided by individual respondents is maintained, and information about the individual respondents cannot be derived from the published results.

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. Interfield 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.

Constant dollars

Dollars of a particular base year, which are adjusted (by inflation or deflation) to show changes in the purchasing power of the dollar (See also current dollars). The base year must always be stated. Note that the terms "uninflated dollars" and "deflated dollars" are used as synonyms for "constant dollars" in government publications.

Continuous variable

A numeric variable which can assume an infinite number of real values. For example, age, distance and temperature are considered continuous variables because an individual can walk 3.642531...km.

Correlation coefficient

A measure showing to which extent two variables vary in an interdependent way.

Cumulative frequency

Determines the number of observations that lie below a particular value. It is calculated by adding each frequency to the sum of its predecessor in a stem and leaf plot and frequency distribution table.

Cumulative percentage

Calculated by dividing the cumulative frequency by the number of observations and multiplying it by 100. The last value will always be equal to 100%. This allows for easier comparison of the data.

Current account

The current account covers transactions on goods, services, investment income and current transfers. Transactions in exports and interest income are examples of receipts, while imports and interest expense are payments. The balance from these transactions determines if Canada's current account is in surplus or deficit.
Source: ARCHIVED - *The Daily*-Friday, May 28, 2004

Current dollars

Dollars which express the cost of items in terms of the year in which the expenditure occurs. Note that "current dollars" are also known as "budget-year dollars" or "inflated dollars", as opposed to "deflated" or "constant dollars".

D

Data

Facts or figures from which conclusions can be drawn.

Database

An organized and sorted list of facts or information; usually generated by a computer.

Data capture

The process of putting responses into a machine-readable medium.

Data coding

Raw data entered into a computer may need to be coded. This is done by labeling each of the data items with an abbreviated code (usually numerical), to make the manipulation of the data easier.

Data editing

A process that ensures survey data is accurate, complete and consistent. (See also: Imputation).

Data input

A process (e.g., scanning, paper, magnetic tapes, cards, etc.) used to enter data.

Data item

The smallest piece of information that can be obtained from a survey or Census.

Data processing

A process that converts raw data into machine readable form, then sorts, edits, manipulates and presents the data in order to create information.

Data quality

A degree or level of confidence that the data and statistical information are "fit for use". The particular issues of quality or fitness for use that must be addressed by Statistics Canada can be summarized as relevance, accuracy, timeliness, accessibility, interpretability and coherence.

Data set

Any grouping of data which has a common theme or similar attributes.

Data storage

The capacity of a computer to store information, as well as the components of the computer in which such information is stored (i.e., magnetic tape, diskette, CD-ROM, etc.).

Decennial

An event recurring every ten years.

Decennial census (Canada)

Censuses are held at the beginning of each decade, in years ending with the number 1. (See also: Quinquennial census)

Discrete variable

A numeric variable that takes only a finite number of real values (e.g., X can equal only 1, 3, 5 and 1,000).

Dissolved census subdivision

The boundaries and names of census subdivisions can change from one census to the next because of annexations, dissolutions and incorporations. These changes can result in the "dissolution" of various census subdivisions. A dissolved census subdivision is a community that existed on January 1, 1996, but which no longer existed on January 1, 2001, the 2001 Census geographic reference date. The concept of "Census Subdivision - Previous Census" has been established to provide a means of tabulating current census data for census subdivisions as they were delineated for the previous (1996) census. A "best fit" linkage was established between blocks for the 2001 Census and census subdivisions for the 1996 Census. This linkage ensures that data from the current census can be tabulated for the communities from the previous census.

Dispersion

Describes how much the observations vary around the central tendency.

Dot graph

A two dimensional diagram that indicates two variables as a series of dots, mainly used to show the correlation between the two variables. Example of a dot graph

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 item has only been recorded once. A duplication edit also checks to ensure that the respondent does not appear on the survey universe more than once, especially if there has been a name change. Finally, it ensures that the data have been entered into the system only once.

E

Editing

See data editing.

Enumeration area (EA)

The geographical area canvassed by one interviewer in the Canadian Census of Population. EAs are units which are well defined and readily identifiable on maps, but are unique to a particular Census.

Estimation

Drawing larger conclusions from a sample to predict some characteristic or trend for the whole population.

Estimator

Uses information from other questions or from other answers (from the current cycle or a previous cycle), and through mathematical operations, derives a plausible value for the missing or incorrect field.

Ethnic origin

Refers to the ethnic or cultural group(s) to which the respondent's ancestors belong. Ethnic or cultural origin refers to the ethnic "roots" or ancestral background of the population and should not be confused with citizenship or nationality.

Exclusive

When the occurrence of one event automatically excludes the possibility of another event occurring at the same time, the results are exclusive. For example, round and square are exclusive terms since an object cannot be both at once.

Exhaustive

When a set of events comprises all possible occurrences of a reference set the results are exhaustive. For example, the list of age groups 0 to 19, 20 to 34, 35 to 59 and 60 years and over is exhaustive because it covers the whole spectrum of possible ages for members of the population.

F

Focus group

An interviewing technique whereby respondents are interviewed in a group setting. It is used to stimulate the respondents to talk freely, encourage the free expression of ideas or explore attitudes and feelings about a subject. It is often used to guide the design of a questionnaire based on the respondent's reaction to the subject matter and the issues raised during the discussion. It is also referred to as an interview or group discussion.

Formula

An equation or mathematical rule.

Frame

A list, map, or conceptual specification of the units comprising the survey population from which respondents can be selected. For example, a telephone or city directory, or a list of members of a particular association or group.

Frequency

The number of times an event or item occurs in a data set.

Frequency distribution

A chart or table showing how often each value or range of values of a variable appear in a data set.

Frequency polygon

A graph formed by joining the midpoints of histogram column tops. (See also: histogram).

Frequency table

A table presenting statistical data by putting together the values of a characteristic along with the number of times each value appears in the data set. Example of a frequency table

G

Graph

Data represented in a pictorial form (e.g., bar graph, line graph, circle graph/pie chart, histogram, pictograph, etc.).

Grouped frequency distribution

The relationship between the values of a characteristic and their frequencies when those values are grouped into class intervals.

Grouped variable

A set of data which has been grouped or classified, according to some common qualitative or quantitative characteristics. Example of a grouped variable

H

Haphazard sampling

A sample selection based on convenience or availability.

Histogram

A graph that consists of a series of columns, each having a class interval as its base and frequency of occurrence as its height. Example of a histogram

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.

Home language

The language spoken most often at home by the individual at the time of the Census.

Hot deck

Uses other records as 'donors' in order 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 this 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.

I

Imputation

Replacing either missing or invalid data with accepted data. Normally performed in accordance with predetermined decision rules. It is often combined with data editing.

Index

A mathematical device or number which is used to express the observation (eg., price level, volume of trade, relative amount etc.) of a given period, in comparison with that of a base period. For example, a cost-of-living index.

Industry

A grouping of producers or service-providers assembled on the basis of the homogeneity of their products or services.

Inferential statistics

The statistical methods used for inferring population values from obtained sample values.

Information

Data that have been recorded, classified, organized, related or interpreted within a framework so that meaning emerges.

Input device

A tool such as tape, cards, keyboard, diskette, CD-ROM, light pen, scanner, digital camera, etc., used to input data into a computer.

Interactive capture

Often referred to as *intelligent keying*. Usually, captured data are edited before they are imputed. However, this method combines data capture and data editing in one function. Although interactive capture is slower, it is a very effective approach to use when there is a lot of interdependency between questions. This process requires knowledge of editing procedures, as the errors need to be corrected right away. Interactive capture also reduces the number of documents handled, as the edits are made directly on a computer.

Internet

A network that links computers all over the world by satellite and telephone, connecting users with service networks such as e-mail and the World Wide Web.

Interpretability

Interpretability reflects the ease with which the user may understand, properly use and analyse the data or information. The adequacy of the definitions of concepts, target populations, variables and terminology underlying the data, as well as the information on any limitations of the data, largely determines the degree of interpretability.

Interquartile range

The difference between the upper and lower quartiles (Q3–Q1) of a data set. This range is used as a measure of data spread: spanning 50% of a data set and eliminating the influence of outliers (the highest and lowest quarters of a data set are removed).

Interval

A set of numbers which consists of those that are greater than one fixed number and less than another: it may also include one or both end numbers. For example, the interval 1.5 –> 3 consists of all numbers that are equal to or greater than 1.5 and less than 3. Note that the number 3 is excluded from this interval.

Intranet

A private, internal Internet with the privacy and security of an in-house system. Intranets can be connected to the public Internet via secure gateways, which have "firewalls" to prevent unwanted external access to internal systems.

J

Judgment sampling

A sample chosen on the assumption that personal judgment and expertise can be the basis of selecting units that are typical or representative of the population of interest.

K

Knowledge of official languages

An individual's ability to conduct a conversation in English only, in French only, in both English and French or in neither of the official languages of Canada.

L

Labour force

Refers to the labour market activity of the population 15 years of age and over, excluding institutional residents, in the week containing the 15th day of the month prior to Census Day. Respondents are classified as either employed or unemployed. The remainder of the working–age population is classified as not in the labour force.

Labour force participation rate

Total labour force expressed as a percentage of the population aged 15 and over. The participation rate for a particular group (for example, women aged 25 years and over) is expressed as a percentage of the population for that group.

Line graph

A graph in which successive points representing the value of a variable at selected values of the dependent variable are connected by straight lines (e.g., unemployment rates among youth over the last ten years). Example of a line graph

Linear correlation

A measure of how well data points fit a straight line. When all the points fall on the line it is called a perfect correlation. When the points are scattered all over the graph there is no correlation.

Local area network (LAN)

A communications network that serves users within a confined geographical area. It is made up of servers, workstations, a network operating system and a communications link.

M

Magnetic recordings

Allows for both reading and writing capabilities. This method may be used in areas where data security is important. The largest application for this type of data capture is the PIN number found on automatic bank cards.

Mainframe

A computer with extensive capabilities and resources to which other computers may be connected so that they can share facilities.

Margin of error

Relative figure that may be expressed as a percentage and is calculated using the sampling error of an estimate. It is used to build a confidence interval for that estimate.

Marital status

The state of being legally married (but not separated), in a common-law union, separated (but still legally married), divorced, widowed or never married (single).

Mean

The most common measure of central tendency, the mean is the arithmetic average of a set of numbers.

Median

The value of the middle item when the data are arranged from lowest to highest; a measure of central tendency. If there is an even number of observations, the median is the average of the two middle observations. In raw data, the median is the middle value, the point at which exactly half of the data are above it and half below.

Methodology

A set of research methods and techniques applied to a particular field of study. At Statistics Canada, methodology refers to survey methodology.

Metropolitan area

Statistics Canada has created groupings of municipalities, or Census subdivisions, in order to encompass the area under the influence of a major urban centre. Specific guidelines are used to group municipalities that are closely interconnected due to people working in one municipality and living in another. The resulting geographical units are called Census metropolitan areas.

Miscellaneous edits

Fall in the range of special-reporting arrangements; dynamic edits particular to the survey; correct classification checks; changes to physical addresses, locations and/or contacts; and legibility edits (i.e., making sure the figures or symbols are recognizable and easy to read).

Midrange computer

Covers a very broad range between high-end personal computers and mainframes. Formerly called "minicomputers", which used dumb terminals connected to centralized systems, most midrange computers today function as servers in a client/server configuration.

Mode

The observation that occurs most frequently in a data set; a measure of central tendency.

Mother tongue

The first language learned at home during childhood which is still understood by the individual at the time of the Census.

Multi-purpose surveys

Survey objectives call for the measurement of many characteristics. For example, a survey on farm expenditures might want to determine more than the overall costs of running a farm. One might want to discover the cost of farm equipment, wages, loans, seeds, feed, etc.

Multi-stage sampling

The process of selecting a sample in two or more successive stages. It involves a hierarchy of different types of units. Each "first-stage" unit is potentially divisible into "second-stage" units and so on.

N

Negative correlation

In a negative correlation, the two variables tend to go in opposite directions. As one variable increases, the other variable decreases. Therefore, it can also be called an inverse relationship.

Nominal variable

Type of categorical variable that describes a name, label or category with no natural order. For example, there is no natural order in listing different types of school subjects: "History" does not have to follow "Biology". These subjects can be placed in any order.

Non-probability sampling

A sample selected by a non-probability method. For example, a scheme whereby units are selected purposefully would yield a non-random sample.

Non-random sample

See non-probability sampling.

Non-response

The situation that occurs when information from sampling units is unavailable for one reason or another. For example, the respondent is unavailable, refuses to answer or refuses to take part in the interview.

Non-response errors

Errors occurring due to non-interviews or non-responses to a specific question on a questionnaire.

Non-sampling errors

Errors caused by factors other than sampling. For example, errors in coverage, response errors, non-response errors, faulty questionnaires, interviewer recording errors, processing errors, etc.

Normal distribution

Often just called the bell-curve or bell-shaped curve. Most of the scores in this graph accumulate around the middle. The mean, median and mode are all equal, and the scores at either end of the distribution occur less often. For example, a curve representing the results of an intelligence test would have the most number of people in the middle or around the 'average' intelligence range. Whereas the number of people decreases as the scores get farther away on either side of the average, giving the curve its shape and name.

Numeric variable

A quantitative variable that describes a numerically measured value (e.g., age or number of people in a household). These variables can be either continuous or discrete.

O

Observation

Data collected for a given variable.

Ogive

The curve on a frequency distribution graph. Note that not all distribution curves have the ogive form. It is therefore better to confine the term to the normal or nearly-normal distribution.

Optical character readers

Or bar-code scanners, are able to recognize alpha or numeric characters. These readers scan lines and translate them into the program.

Ordinal variable

A type of categorical variable: an ordinal variable is one that has a natural ordering of its possible values, but the distances between the values are undefined. Ordinal variables usually have categorical scales. For example, when asking people to choose between Excellent, Good, Fair and Poor to rate something, the answer is only a category but there is a natural ordering in those categories.

Outliers

In a set of data, a value so far removed from other values in the distribution that its presence cannot be attributed to the random combination of chance causes.

P

Parameter

Parameters are unknown, quantitative measures (e.g., total revenue, mean revenue, total yield or number of unemployed people) for the entire population or for specified domains which are of interest to the investigator.

Percentage frequency

The frequency of each value or class interval expressed as a percentage of the total number of observations. Derived by multiplying each of the relative frequency values by 100.

Percentiles

The proportion of values in a distribution that a specific value is greater than or equal to. For example, if you received a mark of 95% on a math test and this mark was greater than or equal to the marks of 88% of students then you would be in the 88th percentile.

Personal computer

A general-purpose, single-user microcomputer designed to be operated by one person at a time.

Pictograph

A chart giving statistics in pictorial form. For example, using a dollar in increasing sizes to represent the increase in the purchasing power over time. Example of a pictograph

Pie chart (Circle graph)

A circular chart that provides a visual concept of a whole (100% = 360 degrees). The pie is divided into slices, each corresponding to a category of the variable represented (e.g. each age group). The size of the slices is proportional to the percentage of the corresponding category. Example of a pie chart

Population

The complete group of units to which survey results are to apply. (These units may be persons, animals, objects, businesses, trips, etc.)

Population pyramids

See age-sex pyramid.

Positive correlation

In a positive correlation, the two variables tend to move in the same direction. When one variable increases, the other variable also increases.

Precision

Precision is a measure of similarity. The same surveys conducted more than once should have the same or similar results. The closer the results from each repetition of the survey, the more precise they are.

Probability sampling

A sampling method in which every member of the population has a chance of being selected. Also called random sampling, because of the random way of selecting individuals to ensure an unbiased representation of the whole population.

Processing errors

Errors that occur during any of the processes performed in transferring data from questionnaires, control sheets, etc., into sets of tabulations and estimates.

Programming

Process of producing a set of instructions to make a computer perform a particular activity.

Pyramids

See age-sex pyramid.

Q

Quality control

The set of operations required to ensure that error levels, introduced as a result of a survey operation, are controlled within specified levels.

Quartiles

In order to determine the interquartile range, a data set is divided into four equal parts. Each separating value is called a quartile (the first, the second, etc.). The second quartile is also known as the median.

Questionnaire

A series of questions designed to elicit information on one or more topics from a respondent.

Quinquennial Census (Canada)

Used to describe Censuses taken at mid-decade, in years ending in the number 6. For example, Statistics Canada conducted quinquennial Censuses in 1976, 1986, 1996. (See also: decennial)

Quota sampling

A procedure where the number of respondents in each of several categories is specified in advance and the final selection of respondents is left to the interviewer who proceeds until the quota for each category is filled.

R

Random error

The errors that are unpredictable in an estimate. These errors tend to cancel out in a large sample, as opposed to systematic errors that keep adding up because they all go in the same direction.

Random rounding

A method whereby all figures in a tabulation, including totals, are randomly rounded (either up or down) to a multiple of "5" or in some cases "10". This technique provides protection against direct, residual or negative disclosure of the actual data, while preserving the usefulness of the data to the greatest extent possible.

Random sampling

See probability sampling.

Range

The full distance over which results vary along a number line. The exclusive range is the difference between the largest and smallest results in a data set, and the inclusive range is the difference between the upper real limit of the highest interval and the lower real limit of the lowest interval.

Range edits

Are similar to validity edits in that they look at one field at a time. The purpose of this type of edit is to ensure that the values, ratios and calculations fall within the pre-established limits.

Ratio

A proportional relationship between two different numbers or quantities, or in mathematics a quotient of two numbers or expressions, arrived at by dividing one by the other.

Raw data

Information that has not yet been organized, formatted, or analysed.

Regression

A statistical method which tries to predict the value of a characteristic by studying its relationship with one or more other characteristics. This relationship is expressed through the means of a regression equation. (See also regression model).

Regression equation

An equation whereby one unknown variable can be predicted using the given value of one or more other variables. For example, the equation $Y = a + bX$ provides the estimated value for Y when the value for X is known. (See also regression and regression model).

Regression model

A statistical model used to depict the relationship of a dependent variable to one or more independent variables. These models have a wide variety of forms and degrees of complexity. (See also regression and regression equation).

Relative frequency

The frequency (expressed as a proportion of a whole) of each value or class interval observed in a data set for a particular variable. Calculated by dividing the frequency by the number of observations.

Relevance

The relevance of data or of statistical information is a quantitative assessment of the value contributed by these data. Value is characterized by the degree to which the data or information serve the purposes for which they were produced and sought out by users. Value is further characterized by the merit of these purposes in terms of the mandate of the agency, legislated requirements and the opportunity cost to produce the data or information.

Response error

The difference between the true answer to a question and the respondent's answer. It may be caused by the respondent, the interviewer, the questionnaire, the survey procedure or the interaction between the respondent and the interviewer.

S

s

The mathematical symbol for standard deviation.

s^2

The mathematical symbol for variance.

Sample design

A set of specifications that describe population, frame, survey units, sample size, sample selection and estimation method in detail.

Sampling fluctuation

The extent to which a statistic takes on different values with different samples. That is, it refers to how much the statistic's value fluctuates from sample to sample.

Sample survey

A collection of information from only part of a population.

Sampling error

An error which arises because the data are collected from a part, rather than the whole of the population. It is usually measurable from the sample data in the case of probability sampling.

Sampling variation

The variation shown by different samples of the same size from the same population.

Scale

A graded line divided into successive values, which may be graphical, descriptive or numerical, used in reporting assessments. For graphs the scale is the subdivision of each axis. The scale may be numerical or categorical. (See also: sample graph.)

School attendance

Refers to either full-time or part-time (day or evening) attendance at school, college or university during the eight-month period between September and May. Attendance is counted only for courses which could be used as credit towards a certificate, diploma or degree.

Seasonal adjustment

A statistical technique used to remove the effect of normal seasonal fluctuations in data so underlying trends become more evident. Economic statistics, which are subject to seasonal influence, are sometimes presented with the seasonal influence removed. The calculated effect of the seasons has been eliminated from the data.

Semi-quartile range

Computed as one half the difference between the 75^{th} percentile (Q3) and the 25^{th} percentile (Q_1). The formula is $(Q_3–Q_1)÷2$. Since half of the values lie between Q_3 and Q_1, the semi-quartile range is one half the distance needed to cover one half of the values.

Σ
(sigma)

A mathematical symbol for adding the values to which it applies.

Simple random sampling

A basic probability selection scheme in which each sample has an equal chance of being selected.

Skewed

Asymmetric distribution of the data values. The values on one side of the distribution curve tend to extend further from the "middle" than the values on the other side.

Software

There are two types of software—system software, which controls the operation of the computer, (e.g., Windows and DOS) and application software (e.g., Word, EXCEL, MS ACCESS and Lotus).

Standard deviation

The square root of variance, standard deviation, measures the spread or dispersion around the mean of a data set. It is the most widely-used measure of spread.

Statistic

A function that will produce a numerical figure whose value may vary with different outcomes of an experiment (e.g. with different samples). For example, the mean or median of a sample, etc.

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.

Statistics

A type of information obtained through mathematical operations on numerical data.

Statistics: Power from Data!

A product from Statistics Canada that will assist readers in getting the most from statistics.

Statistics, the study of

Construed as singular it is the field of study that collects and arranges numerical facts or data, that relate to human affairs or natural phenomena.

Stem and leaf plot

A semi-graphical method used to represent numerical data, in which the first (leftmost) digit of each data value is a stem and the rest of the digits of the number are the leaves. A stem and leaf plot shows all the data values from a sample set.

Stratified sampling

A sampling procedure in which the population is divided into homogeneous subgroups or strata and the selection of samples is done independently in each stratum.

Substitution

Relies on the availability of comparable data. Imputed data can be extracted from the respondent's record from a previous cycle of the survey, or the imputed data can be taken from the respondent's alternative source file (e.g. administrative files or other survey files for the same respondent). This is often difficult to do because, in many cases, there is no other information available than the information provided in the current survey.

Survey

The collection of information about characteristics of interest from some or all units of a population, using well-defined concepts, methods and procedures, and the compilation of such information into a useful summary form.

Survey units

For the purpose of sample selection, the population should be divisible into a finite number of distinct, non-overlapping and identifiable units, so that each member of the population belongs to only one survey sampling unit, to be surveyed only once.

Systematic sampling

The selection of units from a list using a selection interval (k) so that every k'th element on the list, following a random start between 1 and k, is included in the sample. For example, if k were to equal 6, and the random start were 2, then the sample would be 2, 8, 14, 20, etc.

Systems analysis

The process of breaking down a data processing problem into functional components to determine the best method of handling the problem.

T

Tally chart

Used to record data from an experiment, count the occurrences of an event and develop frequency distribution tables.

Timeliness

Timeliness of information reflects the length of time between the information's availability and the event or phenomenon it describes. Timeliness must be considered in the context of the time period that permits the information to be of value and still be acted upon. Typically, timeliness can affect the reliability of the information.

Tree diagram

A branching diagram that shows all possible combinations or outcomes.

- -

Ungrouped frequency distribution

A frequency distribution of numerical data where the raw data is not grouped.

Ungrouped variable

A set of data which has not been grouped or classified but rather a listing of individual observed values (e.g., single years of age).

V

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 reporting time is within the specified limits.

Variable

A characteristic that may assume more than one set of values to which a numerical measure can be assigned (e.g., income, age and weight).

Variance

A measure of spread, calculated as the average squared deviation of each number from the mean of a data set. The term variance also refers to the sampling variation.

Vital statistics

Statistics relating to births, deaths, marriages, health, and disease.

Volunteer Sampling

Samples that consist of people who have volunteered their services, knowing that the process will be lengthy or demanding, and even perhaps unpleasant. These volunteers have often been found to have favourable or a least neutral attitudes, whereas the general population tend to hold a wider range of attitudes on topics of interest.

W

Wide area network (WAN)

A communications network that covers a wide geographical area, such as a province or country. It is different from a Local area network (LAN) contained within a building or complex, and the Metropolitan area network (MAN) which generally covers a city or suburb.

X

$\bar{x}$

Mathematical symbol for the mean of a data set.

X-axis

The horizontal number line on a graph (The Cartesian co-ordinate plane.)

Y

Y-axis

The vertical number line on the (Cartesian co-ordinate plane.)