138-EZ12-1377
Ver. 7

QAB

Quantitative Analysis for Business
—Edition 3.0—

Quantitative Analysis for Business
—Edition 3.0— Contents

Introduction

Managing a modern business enterprise has become an enormously complex undertaking. More and more businesses have been turning to quantitative techniques and analysis as a feasible means for solving their business problems. What points should we bear in mind when planning, starting up, and operating a business on a day-to-day basis? Certainly, invisible factors such as the comfort, passion, and enthusiasm of our customers are important, but there is hardly any need to reiterate the importance of things that can be measured such as money and time.

This course is an introduction of the application of statistics and mathematics to business problems. In order to reach people of all abilities, the course has minimized the amount of mathematical training required.

To start with, we need to understand the meaning of numbers. For example, the general public may have trouble working out the significance of a statement like "The rental fee for a vacant first-floor shop on a Y-meter wide road opposite X station is 100,000 yen per tsubo (about 3.3 m2)," or "The rejection rate for W products manufactured at Z factory is 0.05%." However, except for people who work in these specific industries, most of us cannot make sudden inferences from those figures.

Nonetheless, business figures are not always black and white, neither are they always easy to evaluate. For example, let's assume that the sales forecast for the current period was 10 billion yen at the start of the period, but sales now look set to reach 10.2 billion yen. This means that the figures will be adjusted upward by 2% compared to the forecast, but should we be satisfied with this, or could we have earned even more? Simply staring at the difference of 0.2 billion yen will not reveal the answer.

Here, we need analysis that draws appropriate comparisons, or breaks down the numbers and looks at each factor.

It is essential for anyone involved in business and aspiring to work in management to have basic knowledge of the methods and concepts of quantitative analysis, and to be conversant with the approaches to quantitative analysis.

Unit 1: The Significance of Quantitative Analysis

Unit 1-1: What Is Quantitative Analysis?

What is the essence of quantitative analysis and why is it required?

In essence, results are demanded on a daily basis in business. Business can also be summed up as which actions must be taken in order to get the required results. Sometimes, this is referred to as problem-solving, or formulating strategy and tactics, but it is none other than the question of how to efficiently put together factors (methods) that have a strong causal relationship to the preferable result (objectives).

Actually, since ancient times, we have been familiar with interpreting change through causal storylines in myths, for example.

However, in actual business, multiple factors are intertwined in complex ways and looking at the surface alone does not enable us to properly perceive the causal relationships. Somehow or other, our attention is held by things that catch our eye, or things we are familiar with, and our decisions tend to be hardwired. As a result, weak causal relationships, or something we would have to refer to as fortune telling, which may or may not come true, make us liable to take wrong action.

In situations as complex as these, the objective of analysis is to properly clarify causation in the simplest possible terms and, if necessary, to condense the details. To do so, we need to unravel the relationships between the causal factors (input) that influence output (the output-input mechanism) once we have developed a correct understanding of the results that are subject to analysis (output analysis).

Analysis consists of two types: quantitative and qualitative. Quantitative analysis is based on data that is expressed numerically such as 5 kilograms, 10 meters, 1,000 items, 800 times or 30%. Qualitative analysis is based on information expressed in phrases or statements such as a lot/a little, or easy to do/hard to do. Quantitative analysis, which is the main subject of this course, refers to the process of evaluating and interpreting situations expressed through numbers.

Even though this course focuses on quantitative analysis, an adequate analysis requires a fair balance between the qualitative and the quantitative.

Unit 1-2: Why Do We Need Quantitative Analysis?

The real value of analysis is when it reveals realities that are not noticeable on the surface. With proper training in quantitative analysis, business leaders can select the most appropriate approach and specific quantitative tools for particular types of problems. What are the advantages of performing quantitative analysis?

- Improving decision-making skills
  - Ability to use numbers as a common language to make evaluations instead relying on individual intuition and experience.
  - Ability to use the common language of numbers reduces the risk of misunderstanding or misinterpretation.

- Add more dynamism to decision-making and execution
  - Ability to allocate limited resources to important matters by prioritizing the alternatives.
  - Ability to speed up the decision-making process once the criteria for making decisions have been clarified.

On the other hand, quantitative analysis also has the following limitations:

- May lead to confusion, biased interpretations, or making hasty decisions
  - Overconfidence in "numbers" without fully examining their nature or background (disregarding prerequisites such as data collection, processing).
  - Overlooks human relations and other data that is difficult to quantify because of excess focus on rational numbers-based analysis.

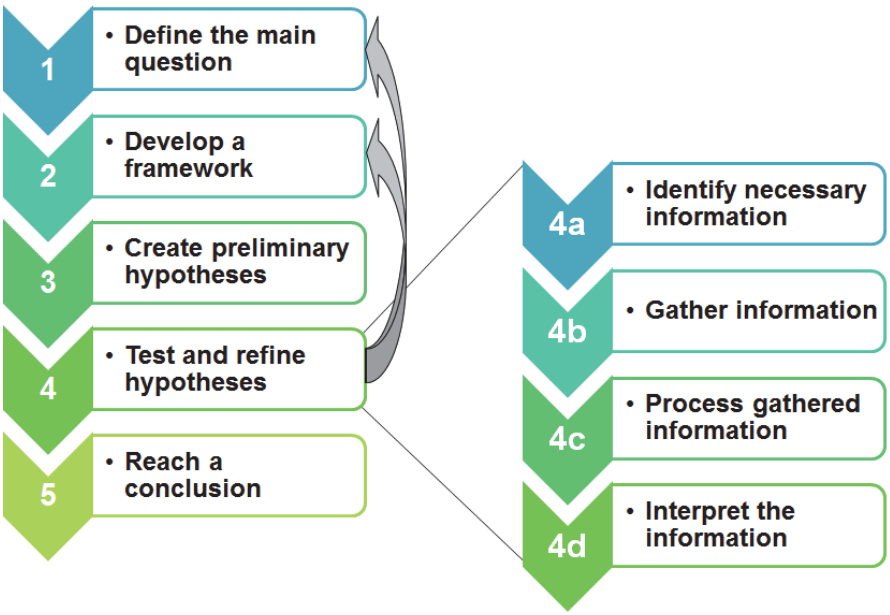
- Numbers assume a life of their own
  - Only the portions that conveniently support a claim are communicated and possibly exaggerated.
  - Continued use and circulation misunderstood and misinterpreted information.

Even with the limitations mentioned above, it is an extremely significant task to master quantitative analysis which can reveal the true nature of things.

For example, a certain market is expected to see slow growth in the future. Let's also assume that by surveying the market and breaking it down into detailed factors (analysis), we find that the market consists of segment A, which is growing rapidly, and segment B, which is declining. You can now decide to focus your actions on segment A, but without this analysis you might have taken sweeping measures aimed at the market as a whole. This is an example of finding out that reality can differ from superficial facts.

Unit 1-3: The Quantitative Analysis Process

The following diagram shows the basic process of quantitative analysis. It is important to first define what the main issue is in the form of a question.



In practice, the process is not a linear sequence of steps, but cyclical and based on feedback. Often, another hypothesis comes to mind as a result of gathering information[1], or you may re-gather information according to another new hypothesis after presenting your interpretation. In some cases, instead of hypothesis, the main question gets redefined.

As further described in Unit 2, these steps are greatly influenced by the likelihood of the hypothesis (what we want to say with this, what we can say, and the provisional messages).

With ill-defined data collection and ill-defined analysis, you will only get ill-defined results. Humans often have a tendency to lean toward analysis based on data without keeping the hypothesis in mind. As much as possible, this course will deal with analysis that emphasizes the steps unfolding from the main question and the hypothesis.

Even though this is quantitative analysis, data processing and calculations should not be the focus nor take a large portion of the process.

Of course, data processing and calculation are indispensable, but it is also important to place roughly equal weight on how to properly establish the main question of the analysis, how to deal with the hypothesis, how to collect information, how to process the information, how to interpret the results, and how to present your findings to others.

[1] In some books, there is a clear distinction between raw data and information. But in this textbook, both terms are used synonymously.

Unit 1-4: Avoiding Pitfalls in Quantitative Analysis

Business leaders and managers should never become captive to quantitative analysis and blindly adopt the conclusions as their answers to the problems. The following is a discussion of some general points when performing quantitative analysis.

(1) Keep your objective, premise, accuracy and method consistent.

In quantitative analysis, it is a common pitfall to play around with the numbers and to do analysis for the sake of analysis. In terms of the accuracy and quantity of the analysis, it is important to adapt to the circumstances in line with the objectives and the premises of what the results of the analysis will be used for, and under what conditions.

Adapting accuracy and quantity to the circumstances refers not only to the final output (how to present the analytical results to others), but also to the process (what degree of data to collect, what degree of accuracy to look for, what analytical methods to select).

Take the following comparisons for example; it would not be effective to aim for the same precision and effort.

Prepare briefing materials for the bank by next week	vs	Prepare materials for internal meeting by tomorrow
Develop plans for a new business	vs	Persuade the company president
Make a decision on an investment of 1 billion yen	vs	Seek approval for expenditure of 5,000 yen

(2) Make appropriate standard for comparisons (apples to apples).

Analysis cannot be conducted without comparison (analysis is comparison). When interpreting the implications of quantitative data, it is important to always have appropriate standards for comparisons, or compare apples to apples (not apples to oranges). When you draw comparisons, such as large or small, you start to interpret the data.

However, since this process in many cases is performed unconsciously, you often come across situations where the analyst does not follow any standard of comparison. Large or small compared to what?

What to choose as a standard for comparison, and whether the standard is appropriate or not, is extremely important for the decision-making process.

(3) Visualize your analysis

Human observe the world with their eyes, information processing heavily depends on our visual senses. Consequently, always visualize every step of the process, from analysis to interpretation to communication. Visualization not only involves looking at printouts of numbers, but also analyzing with our eyes by turning those numbers into graphs and charts.

Visualization is not only useful for effectively communicating messages to others, it is also effective in expanding your ideas, such as discovering new perspective and interpretations as you progress through the analysis.

Below, we compare presentations before and after visualizing the data. Let's think about what tactics have been used, and what effect they have on the person viewing the results

[Case 1] Company A and Company B are rivals in the same market. Both companies surveyed consumers to find out why consumers are not buying their products.

<Before modification>

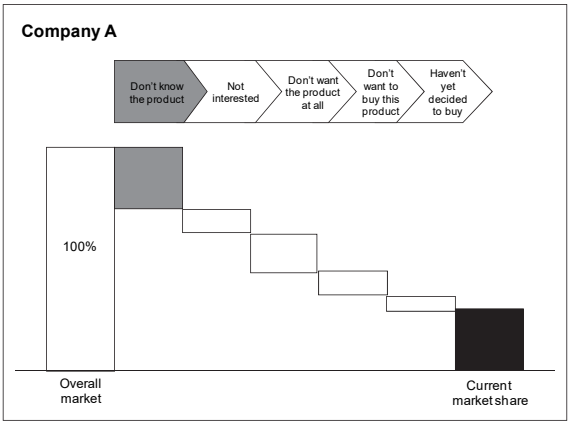
Company A

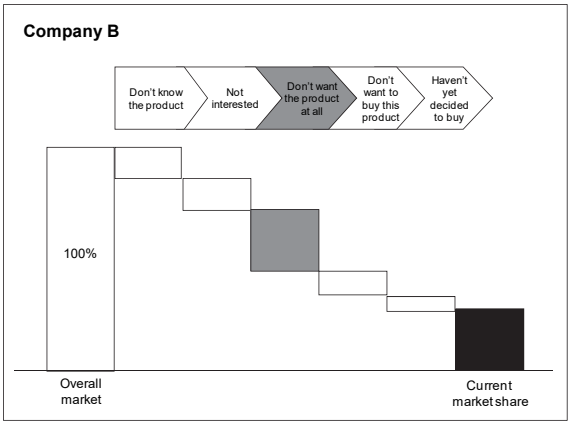
Current market share	30%
Reasons for not buying the product	
Don't know the product	29%
Not interested	10%
Don't want the product at all	16%
Don't want to buy this product	11%
Haven't yet decided to buy	5%

Company B

Current market share	30%
Reasons for not buying the product	
Don't know the product	14%
Not interested	12%
Don't want the product at all	29%
Don't want to buy this product	9%
Haven't yet decided to buy	6%

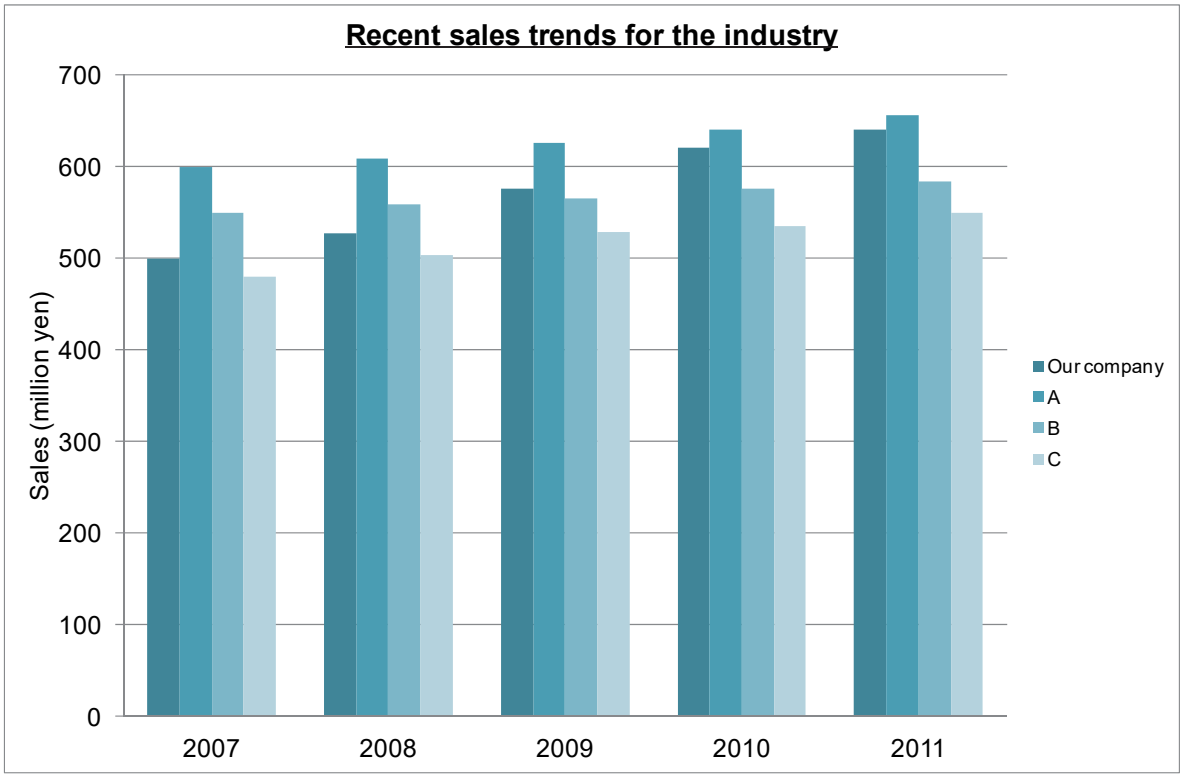
<After modification>



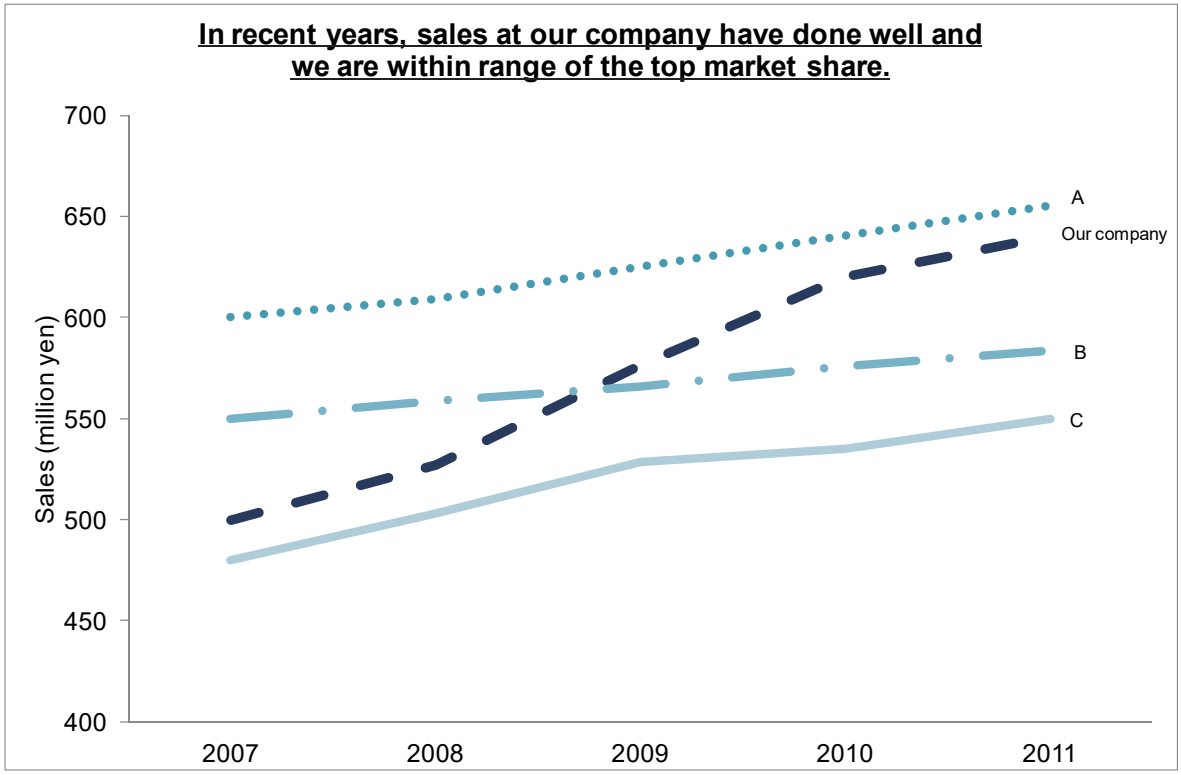


[Case 2] Four companies dominate this industry. A comparison was drawn between their recent sales.

<Before modification>



<After modification>



(4) Draw out messages for decision-making and action

No matter how thorough the analysis, it has no real value in business unless the results can be beneficial for making the final decision. Always ask yourself what the significance of the findings is, and draw out what messages are related to decision-making and action.

Always ask yourself what can be deduced from certain facts and what the implications are. The data and numbers treated in quantitative analysis can be very dull on their own. Therefore, making meaningful interpretations gives you the opportunity to show off your analytical and managerial skills.

Unit 2: Collecting Data

Unit 2-1: What Is Data Collection?

Proper data collection is crucial for accurate analysis. Data collection refers to the act of gathering data not presently on hand, in order to

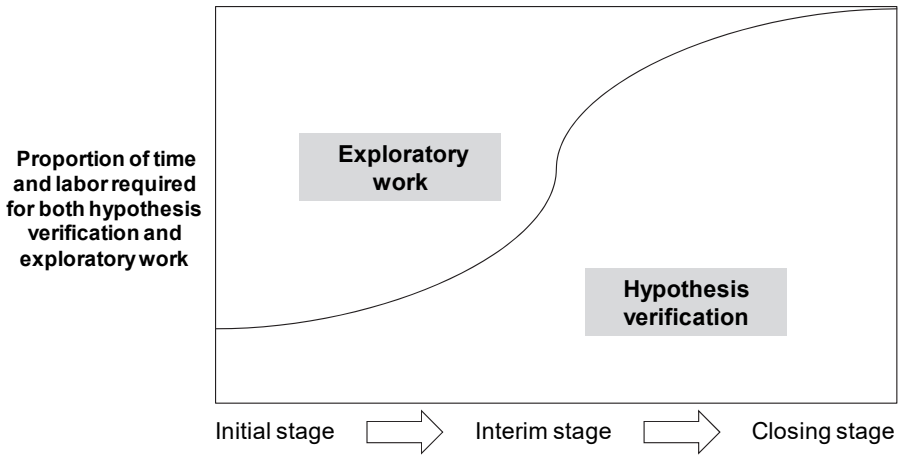
- Investigate a previously formulated hypothesis, or
- Formulate a new hypothesis.

The former is often narrowly focused on gathering the information necessary for testing a hypothesis. In cases where similar surveys have been conducted several times in the past, or where you have come across similar scenarios, it is comparatively simple to establish a hypothesis. At such times, it is possible to carry out focused data collection because you know what kind of information you need to gather to test the hypothesis. Data collection that aims to test a previously formulated hypothesis is referred to as hypothesis testing. For example, in consulting, a storyboard for building hypotheses is created at a fairly early stage, often before starting the actual information gathering. Data collection and analysis is handled in reverse by starting from an image of the final result. The hypothesis is the message headline, and the consultant creates PowerPoint slides with a picture story that already factors in images of the graphs and tables needed to communicate the message before starting to collect the necessary data to produce the message. This is an example of mastering the hypothesis testing approach in a practical setting.

On the other hand, the second type of data collection is often characterized by some degree of comprehensive data collection in order to develop familiarity with the subject of analysis, or to obtain prior knowledge to develop a hypothesis in the first place. It is often used in the early stages of analyzing a field where you have no experience, where you do not know what the important issues are, and where it is even difficult to establish a hypothesis. In such cases, it is important to start by interviewing experts, or by obtaining data that is comparatively easy to find so as to gauge which problems seem important and which hypotheses seem meaningful. Data collection that places the emphasis on producing a hypothesis is referred to as exploratory data collection.

To effectively perform data collection under time constraints, it is important to have some sort of hypothesis when dealing with data collection, even if it is only a “back of a napkin” idea, and to choose the hypothesis testing approach. The hypothesis approach and the exploratory approach are often combined in real business.

The following diagram illustrates the proportion of time and labor required for hypothesis testing and exploratory work according to the stage of the analysis.



Steps of the analysis

In the early stages of the analysis, there is a very high proportion of exploratory data collection, but as the analysis progresses, the proportion of data collected to test the hypothesis increases. The two are by no means mutually exclusive. It is important to put both approaches to their proper use in the analytical stages, rather like the two wheels of a cart.

This figure also suggests that data collection is not done only once (a single round of distributing and collecting questionnaires, and then analyzing the data), but the analysis moves forward while collecting exploratory data and data for hypothesis testing over and over again. This also means that the hypothesis is reviewed and rebuilt and it is often the case that where trial and error and several restarts are involved, the final result is a great piece of analysis.

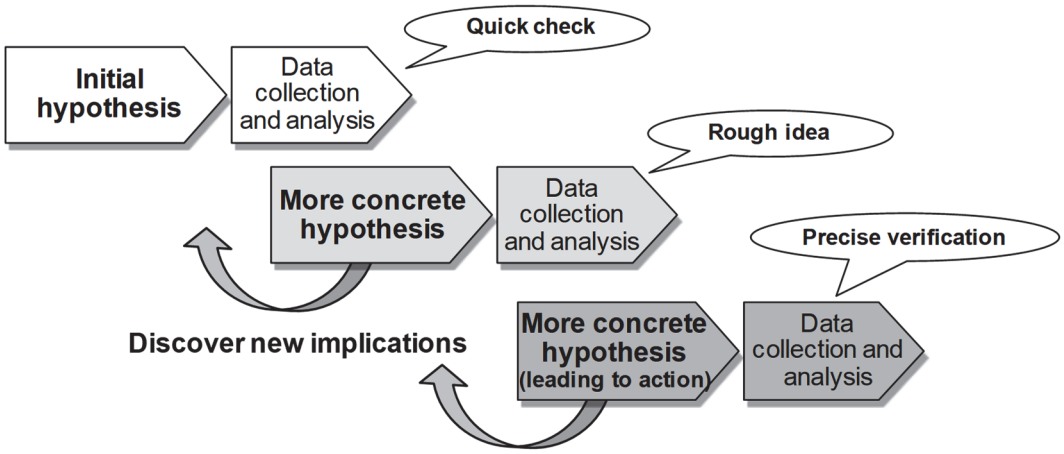
Further, data for collection can be broadly classified as quantitative data (quantitative information) and qualitative data (qualitative information). In Unit 2, we discuss the essence of data collection in general, including both kinds of data.

[Reference]
The word "hypothesis" is routinely used in business. Here, it is defined as a subjective and hypothetical (objectively, we do not know whether it is true or not) answer to an argument or a question (issue).

Briefly stated, there are different levels of hypotheses depending on accuracy (rough or elaborate) and level of abstraction (vague or specific). A rough hypothesis prior to the start of specific analysis or data collection is defined as the initial hypothesis. As stated in Unit 1-4, the degree of accuracy and abstraction of the survey or analysis required for testing varies according to the degree of accuracy and abstraction of the hypothesis.

The initial hypothesis is not the product of a blank slate where there is absolutely no information about the issue, but one that is constructed on the basis of past experience or common knowledge.

Consequently, if you already have an abundance of information and knowledge of the issues in your industry or your occupation, there may be a high degree of accuracy and specificity at the stage of the initial hypothesis. On the other hand, when the industry is completely unfamiliar, or when you are starting a new business, it may be necessary to collect information in order to establish the initial hypothesis. In this case, and unlike data collection to test a hypothesis, you will need to collect extensive information through readily available channels such as reading from general publications or interviewing experts on the issue.

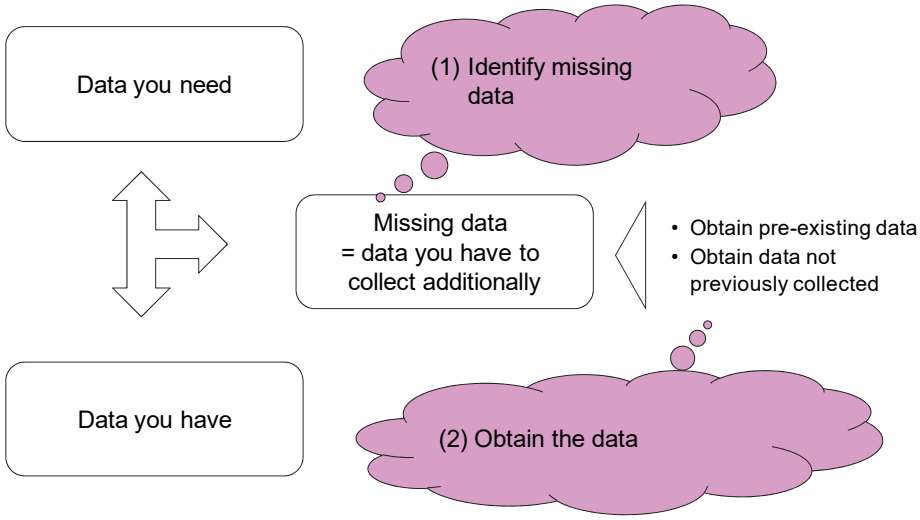

Initial hypothesis
Data collection and analysis
Quick check
More concrete hypothesis
Data collection and analysis
Rough idea
Discover new implications
More concrete hypothesis (leading to action)
Data collection and analysis
Precise verification

Unit 2-2: How to Collect Information

Accurate and efficient data collection is vital for maintaining the integrity of the analysis. By knowing which data to look for and how to obtain the appropriate selection of data, you can not only achieve the business objective but also reduce the likelihood that errors will occur.

Broadly speaking, data collection consists of two steps. To start with, (1) identify missing data and then (2) obtain that data.

<The data collection task>



(1) Identify missing data

Once you have established the hypothesis and understand what data you need for analysis, you need to identify the missing data to prepare your analysis.

Identifying missing data means to shed light on what you already know (data you already have on hand) about the subject of analysis and what else you must know (other data you have to collect).

For example, if you want to test the hypothesis that the majority of users of product A are women in their early 20s, you will at the very least need information on their gender, age and frequency of using product A. Therefore, the data that you should have on hand for a concrete analysis is gender, age and frequency of using product A. If you do not already have this information, you will have to collect the data.

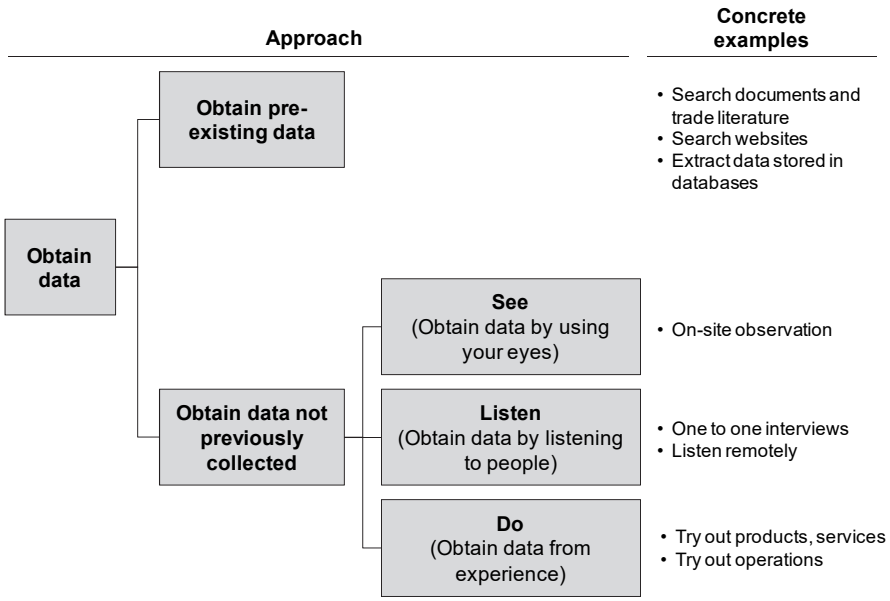
By doing this, you will prevent a situation where you discover that you have spent time and effort collecting additional data when a little coolheaded thinking would have revealed that you already had data on xxx at hand (or could have obtained it easily). Your data collection will also be more efficient because you will be able to focus on collecting the data that you really need to collect.

Before starting the practical work of obtaining data, you need to clarify what you already know (data you have already obtained) and what you do not know (data you actually have to collect).

(2) Obtain the data

Once you have clarified what additional data is needed, you need obtain that data for processing, analyzing and interpretation.

The following outline is one approach to obtaining data.



The following is a detailed description of each step.

- Obtain pre-existing data
  Obtaining pre-existing data means that the data you need to obtain already exists somewhere in the world. (You have not come into contact with it yet, but if you search, you will find it.)

  Specifically, you can use the following methods:

  (Examples)
  - Search documents and trade literature (publications, magazines, newspaper articles, theses, trade journals and trade periodicals, statistical data published by government offices, industry reports published by industry groups or private research institutes)
  - Search websites
  - Extract data stored in databases

  A typical example of obtaining pre-existing data is to look at documents, trade literature and websites where the data has been compiled in some shape or form. Extracting raw data in the form of rows of numbers (data that has not been processed in any way) from database also falls under this approach. Even though the data as such has not been neatly organized, it has already been collected. If you extract it from the database, you will be able to organize the data you need for analysis.

  Generally, this is a common approach in the early stages when you start to obtain data. Compared to the other approach of obtaining data that has not been previously

collected, this is potentially an extremely efficient approach as long as you are skilled at identifying the location of the data. Search tool development has also boosted the authority of this approach. However, to make more effective use of this approach, it is important to create your own lists of what type of data exists (or is likely to exist) where, and not to rely on search tools alone. In particular, every time you complete a round of data collection tasks, it is a good idea to continuously update your list of information sources by noting where you found (or did not find) what types of data.

- Obtain data not previously collected

Pre-existing data does not necessarily provide you with the most appropriate data that you need for your problem. Pre-existing data was often collected for a different purpose, so it is not necessarily suited to your purpose. Consequently, (at the start of data collection) you need to collect and obtain data that has not yet been collected.

Broadly speaking, there are the following three methods.

1. See (Obtain data by using your eyes)
2. Listen (Obtain data by listening to people)
3. Do (Obtain data from experience)

1. See (Obtain data by using your eyes)

This refers to a method of obtaining data by going to the actual field and observing the situation with your own eyes. This includes observing customer purchasing behavior, surveying competitors, and observing operators at your company's factories.

Heading out to a site and seeing the situation for yourself is important because these days it is so easy to collect information from documents and the internet. They say that a picture is worth a thousand words, and by observing a situation for yourself, you may see things in a different light than what is described on paper or from the internet. You may also be able to obtain data that differs from the prevailing opinion, making it easy to identify new opportunities.

2. Listen (Obtain data by listening to people)

Obtaining data by listening to people refers to the method of obtaining data by listening to experts, concerned parties, or the general public.

The following are the main examples.

(Examples)

- Listening to someone in person
  - One-on-one interviews
  - Group interviews
  - On-site test (known as "central location tests")
  - Conducting the interview at one's home, etc.

- Listen remotely
  - Mail-in questionnaires
  - Telephone questionnaires
  - Online questionnaires & interviews, etc.

Although we call them interviews and questionnaires, they are not always formal sessions featuring an exchange of prepared questions and answers. In business, for example, it is also important to first ask people who seem well informed. This demonstrates authority, particularly when you are familiarizing yourself with a field where you have no previous experience.

Generally, when obtaining data by talking to people, the pitfalls include inappropriate selection of interviewees, inappropriate sampling, or inappropriate questions (details to follow). You should pay particular attention to sampling and the phrasing of the questions.

3. Do (Obtain data from experience)

Obtaining data from experience refers to data collection based on actual experience while interacting with the subject of analysis as an interested party. Obtaining data about operational bottlenecks by working in product assembly at a factory is one example. Another example of obtaining data from experience is to use the products and services of the competition to obtain material on how you can improve your own products and services.

By getting out in the field, trying things out on your own and gaining first-hand knowledge, it is possible to collect valuable data rooted in real experiences, which are difficult to express in words and numbers.

Unit 2-3: Avoiding Pitfalls in Data Collection

In this section, we introduce some important points to keep in mind when collecting data focused on listening to people, because listening is one of the most effective data collection methods in business.

(1) Inappropriate sampling

- Insufficient data and sample size
  You need to pay attention to whether the collected data and the sample (a set of data collected from a statistical population defined by a range) reflect the actual situation of the subject for analysis and represent the parent population (the total population from which the sample is taken). If you have little data and few samples, you are likely to overestimate unique data that occur unexpectedly, or the opinions of people with loud voices. Generally speaking, the law of large numbers presumes that the more data you have and the higher the number of samples, the closer you get to the reality of the parent population. In business, however, it may be difficult to collect a sufficient number of samples due to the effort, cost and time involved. If statistically reliable data is required, calculating the standard error rate is one method of finding out how far removed the sample is from the parent population. Above all, though, it is important to acknowledge whether or not the collected data is representative of the subject for analysis and the parent population before analyzing and making judgments.

- Skewed data or samples
  It is essential to confirm whether or not the data or the sample are skewed when checking them against the objective of the analysis. For example, when analyzing factors shared by successful companies, the analysis does not really have any meaning if you only analyze factors from successful companies because it is possible that the success factors are also present at unsuccessful companies. In this case, it is important to collect data on both successful companies and unsuccessful companies to identify the factors that are present in the former, but not in the latter, and to assess the factors that represent the watershed between success and failure. In this way, you have to be careful about any skewing of the data or the sample.

(2) Inappropriate Timing of Data Collection

- Seasonal bias
  Seasonal bias refers to bias in the responses (the data) that is due to the season, or the time when the questions were asked of the respondent, or the period of data collection. For example, to develop new products, a *soba* and *udon* restaurant conducts a questionnaire in which they ask respondents to select their favorite foods from a list. If the survey is done at the height of summer, more respondents will say *zaru soba* or *soumen* while fewer people will select *nabeyaki udon* or similar dishes (which are popular in winter). If the survey is done in the depths of winter, you should see the reverse trend. If the aim is to develop a product that will sell throughout the year, it is advisable to consider the variability of the responses according to season, and to standardize the responses by collecting questionnaires throughout the year.

- The effect of recent events
  This refers to the tendency to overestimate recent events, which is often identified in employee performance evaluations. Even though performance should be evaluated over the whole year, the evaluation is often greatly influenced by events a few months before the end of the year. It is advisable to pay attention to any effect of recent events on the collected data and responses.

- Dramatization effect
  The reverse of the effect of recent events, the dramatization effect places excessive emphasis on experiences in the past, especially dramatic events during one's early years. Before you know it, a story about dealing with customer complaints all night for two days in a row can become overstated into a story of five days in a row, as time passes by. So, it is advisable to pay attention to the differences between childhood sensations and your experiences as an adult.

(3) Inappropriate means for collecting data

- Inappropriate medium for collecting data
  The previously mentioned bias in data and samples occurs if the medium for collecting data is inappropriate. One example is using questionnaires on the Internet. If you conduct an online survey with the question "Do you like to read ebooks?", chances are you will receive a comparatively large number of affirmative responses. However, in this case, it is unlikely that you will have any responses from people who seldom go online. If the objective is to understand general trends of consumers' reading behavior, this collection method is inappropriate. It is advisable to think about what collection methods are appropriate in the context of your objectives.

- The data collector (interviewer, etc.) influences the data
  The attributes of the interviewer may influence the respondents during face-to-face questionnaire surveys. For example, in an internal company survey, the question "Are you satisfied with the company where you work?" may elicit different answers from the same respondent depending on whether the interviewer is an executive or a new, young employee. It is important to confirm whether or not the attributes of the data collector influence the data contained in the answers.

(4) Inappropriate questions

- Ambiguity in the meaning and scope of the words
  When questions contain words that are ambiguous regarding time and scope (for example, sometimes, often, frequently, near here, around location xxx), they can be interpreted in various ways. Even if you ask someone to tell you their annual income, the respondent may give an answer that differs from the original intent of the question, since you have not defined whether you are referring to last year's performance, or the estimate for this year, or whether their bonus was included or not. It is advisable to clarify the scope, term and unit of the question and avoid words or expressions that are ambiguous in terms of meaning or degree.

- Inquiring about several things with one question
  Avoid asking about several things with one question, such as "Do you think this product is good for beauty and health?" or "Would you recommend this product because it is cheap?" In the second case, it is difficult to accurately reflect the opinions of people

who think the product is cheap, but would not recommend it. It is important to ask only one thing with one question.

- Leading questions
  Avoid using expressions or phrases that suggest an answer. For example, if someone is asked, "Do you agree with the amendments to this law that may encourage certain types of crime?", it is very difficult to say yes. If you are asking a series of questions, you also need to pay attention to the order of the questions. For example, if you ask about the intent to purchase after a question about the advantages of a product, you will see a trend toward many affirmative answers. In this case, you need to devise a strategy such as asking about the intent to purchase first, or interspersing several other questions before you ask about it. It is advisable to check if your choice of expression, or the order of the questions, might suggest a particular answer.

- Omitted or overlapping options
  If you want the respondents to choose from pre-existing options, you need to make sure there are no omissions or overlapping options. For example, "Married/Single" may be too simplistic. Instead, consider Married, Single (never been married), Divorced and Widowed.

Unit 3: Analyzing Data

Unit 3-1: The Objective of Analysis

Now that we have collected the data, we can begin analysis. To begin, let's review the purpose of analyzing data in business.

Management is nothing less than the continuous pursuit and implementation of methods to realize the goals that you have set for your business. For example, when Company A, a manufacturing business, announced its global vision, the company's president clearly stated in his message that in terms of earnings structure, the company aims for "the early delivery of a strong revenue base, which allows the company to consistently generate 1 trillion yen in revenue from sales of 7.5 million products and an operating profit margin of 5%, even in a severe business environment and the dollar exchange rate is at 85 yen." To achieve this goal, the company would have to undertake a range of measures including improvements to the earnings structure.

Here, goal-method and causality are inseparable. Causality between goal and method is a requirement, and, if you understand causality, you will be able to think up effective methods for achieving the goal. There is an expression: "strategy is a story." The relationship between goals and methods that we call "strategy" is expressed in the language of stories, which is a stream of dynamic causality. Surely, this is the essence of business. Consequently, in business, the goal of data analysis is in the end to understand any hints of causality that can serve the relationship between goal and method.

Normally, we use the expression "cause and effect" when addressing causality. In the reality of business, causality is much more complex. You may have a chain reaction where the outcome of a particular cause and effect turns into the next cause, so in this course, we refer to anything that is to some degree situated in the causal system as input, and anything that is, by comparison with the input, situated in the effect system, as output.

In actual analysis, we try to understand the following two points:

- Analyze the output (the eventual condition)
- Analyze the mechanism between output and input (causality, logical interaction)

When we analyze output, we determine the characteristics of the output (what the overall situation is like, where the business opportunities are, or where the problems are) before clarifying the mechanism.

When we analyze the mechanism between output and input, we analyze the logical connections between output and input. As already mentioned, the focus is causality, but in business, temporal causality is not the only commonly used logical connection. There is also logical interaction for explaining structures that can be understood through modeling (for example, sales = unit price × number sold), which we will discuss at a later stage. Consequently, in this course, we do not presume to use the term causality, but "mechanism," to understand logical connections, including causality, as well as logical interaction.

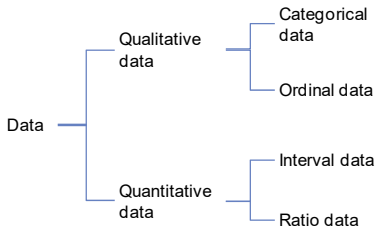
Unless it is processed, data is no more than a series of cut and dried numbers. To gain suggestions for business by means of data analysis, the "perspective" and the "approach" play extremely important roles.

Data analysis = Analyze output + Analyze the mechanism between output and input
Analysis = Perspective × Approach

In this unit, you will learn about perspectives and approaches to analysis.

Unit 3-2: Types of Data

Before moving on to perspectives and approaches, let us review what types of data exist in the first place.



Summary	Permitted operations	Example
Data for distinguishing attributes (numeric conversion for convenience)	Numeric conversion for distinguishing attributes (e.g. 0 men, 1 women), frequency counts only	Gender, occupation, telephone number, blood type
The sequence is meaningful, but the intervals are irregular or unclear.	Size comparisons (<, >)	Customer satisfaction (5: Very satisfied, 4: Satisfied, etc.)
The size of the intervals are meaningful, but ratios have no meaning.	Addition and subtraction (+ and -)	Temperature (Celsius, Fahrenheit), IQ
Data with a natural zero point, such as length or weight, where ratios are meaningful.	All 4 arithmetic operations (+, -, x, /)	Pricing, length, weight, Temperature with an absolute zero

Categorical data is a type of qualitative data that shows the general attributes of the data, which can be numericized for convenience. For ordinal data, the data comes in some sort of sequence, but the size of the intervals between the data can be different or undefined. Interval data is quantitative data where the size of the interval is defined and meaningful. Finally, ratio data has an "absolute 0 point," where having a "zero value" is meaningful. (In Celsius and Fahrenheit, there is no such thing as "no temperature," so temperature is interval data.) Having a zero point allow for ratios to become possible, thus allowing for more statistical possibilities. However, since you can calculate averages, standard deviation, etc. for both ratio data and interval data in quantitative data, in practice there are not so many occasions that you have to be aware of the difference.

The results of questionnaires measuring satisfaction on a 5-point scale (5 Very satisfied, 4 Satisfied, 3 Not sure, 2 Dissatisfied, 1 Very dissatisfied) are, strictly speaking, ordinal data, but in practice, people sometimes turn a blind eye to this and extract averages by treating it as interval data.

Quantitative and qualitative data serve different purposes and provide different results. When are used together, they can present a comprehensive situation.

Unit 3-3: Perspectives on Analysis

1) Five Perspectives on Analyzing Outputs

If you think carefully about perspectives in analysis, you soon realize they can be summed up as standards for comparison (as in when we compare apples to apples, as previously mentioned). The implications can be discovered by aligning the axes for comparison and so that a comparison can be conducted. In the following section, we will learn about five perspectives and focus on what to compare and what the considerations are.

(A) Impact
(B) Gaps
(C) Trends
(D) Dispersion
(E) Patterns

(A) Impact

The first perspective is the size of the impact of the subject for analysis. In short, think about how big an impact the subject for analysis will have on the final results. That is, you need to be aware of the final outcome, which is the problem to be resolved, and the effect of implementing the solution strategies.

For example, suppose a manufacturer of medical equipment with sales in the 100 billion yen range is studying whether or not to enter the booming market for surgical instruments. To consider the size of the impact means to understand whether the scale of the surgical instruments market is 1 billion yen, 10 billion yen or 100 billion yen in terms of size. Certainly, the market may seem attractive from the growth rate perspective since it is growing rapidly. However, if, for argument's sake, the market is doing well, but the scale is 1 billion yen, it may not even be worth investigating a market entry in the first place. The reason is that, by and large, the amount of profit you can draw from that market is limited if the market scale is at most 1 billion yen. Even if you decide to enter the market, the sales targets, the means and the timing of entry will differ depending on whether you are entering a 1 billion yen market, or a 100 billion yen market. When you undertake analysis, it is important that you understand the size of the subject for analysis.

So, why is it important to understand the size of impact in business? The reason is because business resources and time are limited. To most efficiently accomplish the greatest benefit with limited resources and time, you have to avoid spreading your resources too thinly. To put it differently, it is advisable to concentrate investing resources in the area with the most impact on the required result.

Inadvertently, we have a tendency to emphasize the opportunities and problems that come to our attention and think about how we will respond to them. However, there is no guarantee anywhere that the opportunities and problems you notice will have any great impact on the result. Inexperienced leaders, in particular, are obsessed with problem solving and have a tendency to make mountains out of molehills. Consequently, in business, we need levelheaded thinking, such as "How big is this opportunity?" or "We are attempting to solve these problems, but how big will the impact be?" You need to find the areas where the impact will be great, prioritize and take charge. To put it differently, from time to time you need to ignore small problems.

(B) Gaps

By comparing the subject for analysis with something else, we recognize variances between the subject for analysis and the subject for comparison, i.e., what is the same, what is different and how. These variances are referred to as gaps. By thinking about why something is the same or different, we can understand the inherent characteristics of the subject for analysis. In reality, there is an element of comparison incorporated into all analysis; we might even say that analysis is comparison. It is no exaggeration to say that it is only by comparison that we can interpret data.

Business frequently uses analysis that focuses on gaps (increments) by drawing comparisons. For example, variance analysis focuses on gaps by drawing comparisons with set targets, planned values and benchmarks. In problem solving, we compare the ideal with the present situation, and if there is a gap we treat it as a problem and undertake analysis to find out what caused the gap. Comparison is also important when we make inferences about the output/input mechanism and the causality mentioned previously.

In the business context, we often use expressions of comparison such as high and low, or large and small, but there are still cases where it is not clear what kind of comparison brought the gap into view in the first place. Consequently, it is extremely important for decision-making that you clarify the subject for comparison, what you have selected as the subject for comparison, and the axis of comparison, and whether the choices are appropriate.

Assuming a different outlook and imagining yourself as one of the people involved is an effective device for becoming aware of multiple perspectives when considering a subject for comparison. For example, suppose that Japan's GDP grew by a rate of 2% in a certain year. Is this high, or low? In this case, you would try to see this from different standpoints, for example, the government, the opposition parties, mass media or foreign investors.

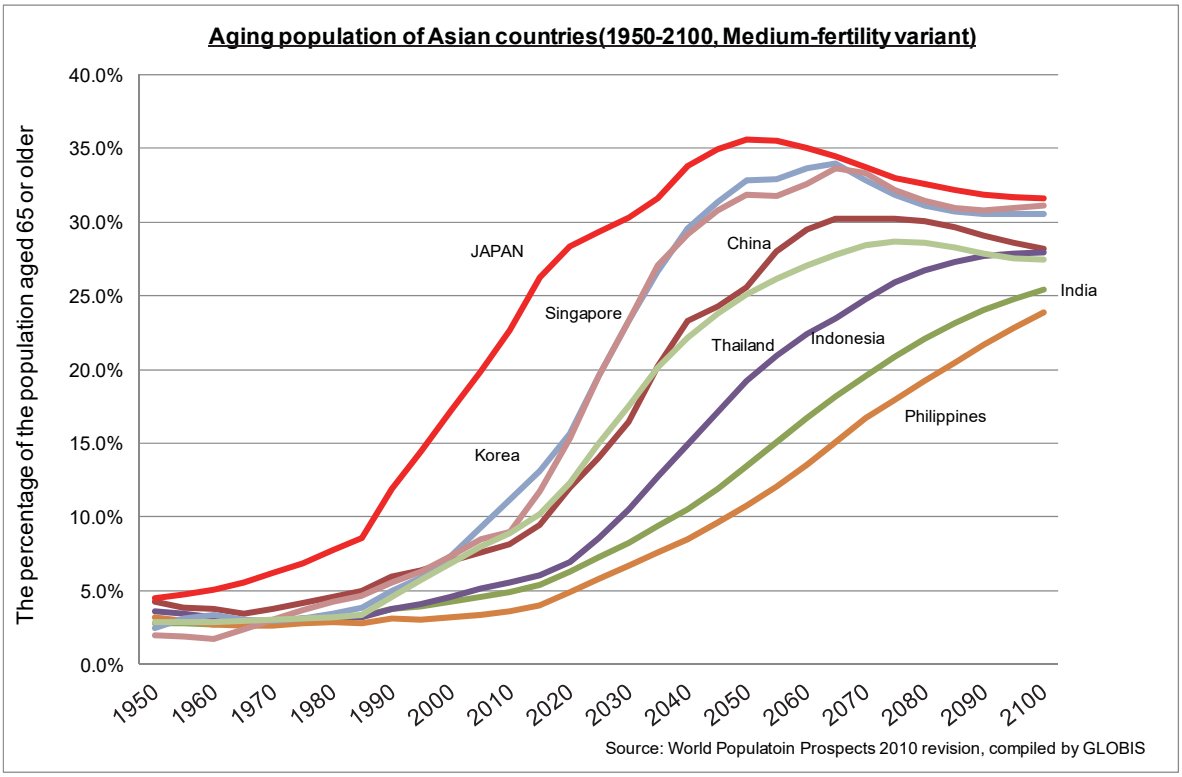
As previously mentioned, it is also effective to visually associate yourself with the context for decision-making in a way that is as realistic as possible. For example, when you consider whether the price you want to set for a new product is expensive or cheap, imagine a situation where you are handling rows of similar products in a store and visualize your surroundings.

(C) Trends

When we look at trends, we pay particular attention to changes on the time axis and, specifically, any trends or deviations from the trends (inflection points etc.). In short, when comparing the past, present and future on the time axis, you consider what kind of trends there are, whether they are increasing or decreasing, and about growth rates. For example, analyses of sales and profit fluctuations, or population changes start by looking at trends.

The graph below projects aging (percentage of population aged 65 or older in the whole population) in Asian countries up to the year 2100. The aging ratio is an indicator of the extent of aging in a country. Generally, a ratio exceeding 7% is considered to be an ageing society, where 14% or higher is an aged society, and a ratio above 21% is referred to as a super-ageing society. The trends in the graph indicate what we already know from discussions in the media: Japan is in the vanguard of ageing in Asia with ageing continuing at full steam until the middle of the 21st century. At the point where the curve peaks, nearly one in three people is aged 65 or older. In fact, however, we also understand that in the 21st century, the outlook is that rapid aging will occur in all Asian countries, not only Japan, and that, therefore, aging is a shared concern in Asia. For example, China is going down the same path of ageing, but roughly 20 years behind Japan. At the start of the Internet age, the method of management that acted swiftly to bring successful business models from overseas to Japan was referred to as a "time machine business," because they were ahead of other domestic business models. Similarly,

when we compare these trends, we understand that Japan's situation and experience as a super-aging society is a time machine for the future of other Asian countries.



(D) Dispersion

Looking at dispersion means to understand the degree of dispersion of the factors that make up the whole, in short, whether the factors are distributed in a specific area (concentration), or distributed evenly across the board. In reality, a lot of things in the world are distributed unevenly and often one part may have a large impact on the whole. This has been a rule of thumb for a long time, for example, in the form of the Pareto Principle (the 20/80 principle). Generally speaking, it is well known in the business world that the top 20% of customers account for 80% of sales.

Since there is a limit on the availability of resources and time in business, it is extremely important to start by dealing with the things that are important, or the things that are highly accessible. Observing particular concentrations is very useful for setting priorities in this way. If we apply the Pareto Principle to problem solving, we will find solutions to 80% of the issues by dealing with the top 20% of the problems.

Using the POS (point of sales) at convenience stores to analyze what items sell well and what items move slowly is certainly a type of analysis that looks at dispersion and concentration in the demand for products. Convenience stores are expected to maximize earnings with limited store space, so they have no choice but to refine product lines. Consequently, we can understand their methods of analysis by how they identify products that do not sell, refine their product lines by cutting products from the stores, or replace them with new products through merchandizing.

However, in business it is not only important to use dispersion to set priorities, but it is also important to remove dispersion and improve homogenization and standardization. For example, for the Japanese way of business after World War II, reducing dispersion in quality management was certainly an overriding factor. Also, when demand fluctuates (with regards to time), how to allocate facilities and staff is an important issue in many fields.

(E) Patterns

Patterns include regularities concealed in the subject for analysis, any diverging singularity, and any inflection point where trends undergo significant change. Below, we will look into each point in more detail.

- Finding patterns (regularities)

  Finding regularities means to identify rules (for example, "when characteristic A is present, xxx happens," or "when characteristic B is present, the result is yyy,"), or the trends and cycles that repeatedly emerge in a given period.

  For example, retail store management constantly attempts to find these kinds of regularities. Stores are required to somehow find regularities in sales ranging from weather conditions (on a sunny day, we sell xx, but if the temperature rises above X degrees, we sell yy) to ingredients used on such-and-such a TV program, or fashion items worn by a particular model. An example of finding repeatedly occurring patterns is when record stores and bookstores try to assess how long sales will remain constant after the release of new DVDs or newly published books, and at which point demand will start to slow.

  Finding these sorts of rules and patterns is referred to as finding regularities.

  What are the advantages of finding regularities for businesses? By finding regularities, it is possible to increase the accuracy of estimates and the reproducibility of courses of action. In the retail example mentioned above, you could grow sales by stocking up on products that are likely to sell, or by being creative with displays within the scope of the regularities you have discovered. At a DVD chain store, it might be easy to grow sales for the whole chain by prominently displaying new titles at the stores until the fourth week from the release, and replacing it with some different new titles from the fifth week (assuming that our analysis has showed that sales slow after the end of the fourth week). If you can lead the competition in the discovery of such regularities and establish a cycle that quickly implements the corresponding measures, you will build up an advantage that is not easily imitated.

  Further, finding regularities includes the legwork for discovering singularities and inflection points. Singularities demonstrate characteristics that differ from the ordinary while inflection points indicate the point where past trends change. Both are recognized because they are based on regularities, which is that "typically, things are more or less like this." Be sure to accept the challenge of searching out regularities in order to find both the singularities and the inflection points discussed below.

- Finding singularities

  Finding singularities means to identify the factors that demonstrate characteristics that differ from rules and patterns. Take the example of the DVD chain store mentioned previously, where sales of new titles often fall after the fourth week, but there is one store where the same trend is not observed. Suppose sales of new titles do not track the downward curve even in the fifth week, but only slow down in the eighth week. We can say that this store is a singularity demonstrating characteristics that differ from the general pattern.

  There are two advantages to focusing on singularities. First, the singularity itself conceals a business opportunity that could not have been predicted in advance, and, by figuring out the mechanism that created the singularity, it is possible to obtain some hints for the business. For example, suppose that in the example of the DVD chain store, a certain store employee creates a promotional POP (point of purchase) and

product displays at the store where the singularity was found, and that this is linked to sustaining the demand for new DVD titles. If you know this, you may be able to grow sales at other stores by using similar solutions. In this way, the singularity may serve to create potential for more business opportunities.

Secondly, the singularity may inspire us to see the subject of analysis from a different angle than in the past. From a particular perspective, it may simply be a singularity, but by searching out different perspectives that also include the singularity, we can gain a more fundamental understanding of the situation. For example, suppose that by separating stores that use POP and displays based on the original ideas of the employee mentioned previously, from stores that do not, we find out that the former stores perform better. Based on this discovery, we may be able to improve sales for the whole chain by researching the store management methods that inspired the ideas of that employee, and by reflecting them across the whole chain.

- Finding inflection points

  Finding the inflection point means to find the point where we start to see rules that differ from the regularity observed in past and where change is rapid.

  For example, suppose that sales of the oral rehydration solutions used to treat heatstroke remain constant if the temperatures never reach 30˚C, but as soon as the temperature exceeds 30˚C, there is a rapid spike in the sales growth. In this example, 30˚C is the inflection point. If you stock the same amount as in the past even on days when the temperature is likely to exceed 30˚C, you will miss an opportunity to make sales. Conversely, if you only perceive that sales increase if it gets hot and stock up even when the temperature is still below 30˚C on the assumption that the temperature will rise, you will end up with unsold items. If you understand this inflection point in advance, you will be able to stock proper quantities. This is the benefit of understanding the inflection point.

Finding patterns in large amounts of data

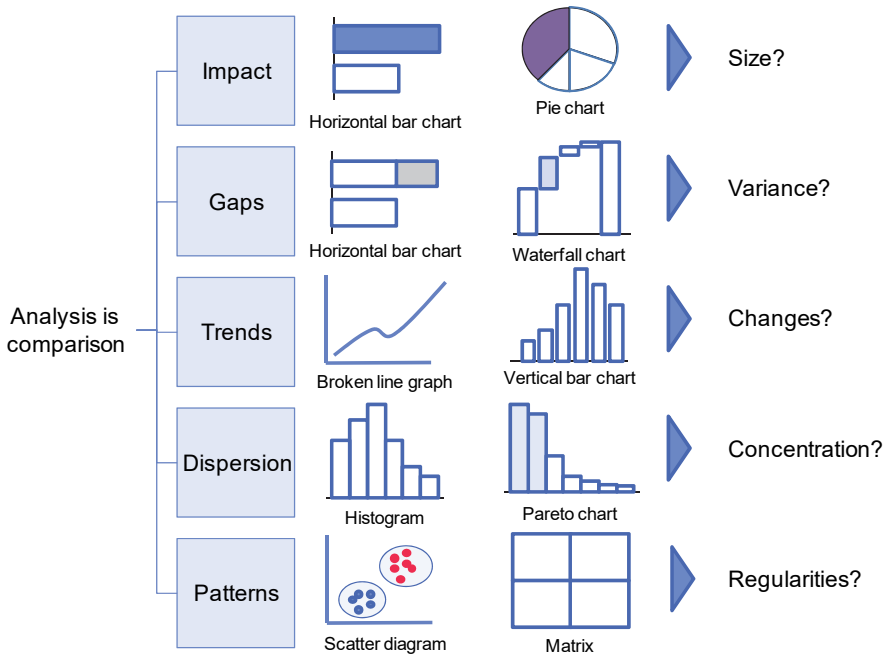
Large amounts of customer data, including credit card purchase history and website browsing history, is now stored at corporations. Due to advances in IT (information technology), corporations analyze these huge amounts of data and by identifying patterns, try to make use of the data to seek out products and services that appeal to consumers and fit their consumption behaviors. One example is the recommendation function at online stores, which uses the correlation coefficient (discussed in Unit 3-4.3). The following example introduces Target, a U.S. company, and tells the story of the predictive power of patterns found by analyzing large amounts of data.

An example at Target (a supermarket chain in the United States)[2]

One day, an angry father stormed into a Target store in Minnesota and started shouting, "What do you mean by sending coupons for maternity wear to my high school daughter!!" The store manager apologized on the spot, but when the manager telephoned the father a few days later, he apologized in return, saying that he had had a good talk with his daughter and that she was expecting a baby in August.

Why was Target able to predict a pregnancy that not even the family knew about? The answer lies in purchasing patterns. The company had built a predictive model for pregnancy by extracting shared purchasing patterns from the purchasing data of customers who had had children in the past. For example, if a customer purchases large amounts of fragrance-free lotion, vitamins, zinc or other supplements, as well as large quantities of fragrance-free soap, the expected delivery date is near. Families with newborns have a tendency to buy things in large quantities, which is a big business opportunity for supermarkets. By detecting purchasing patterns linked to delivery, the store was able to efficiently reach a potential customer.

Here is a summary of the five points and examples of charts used for visualizing the data.

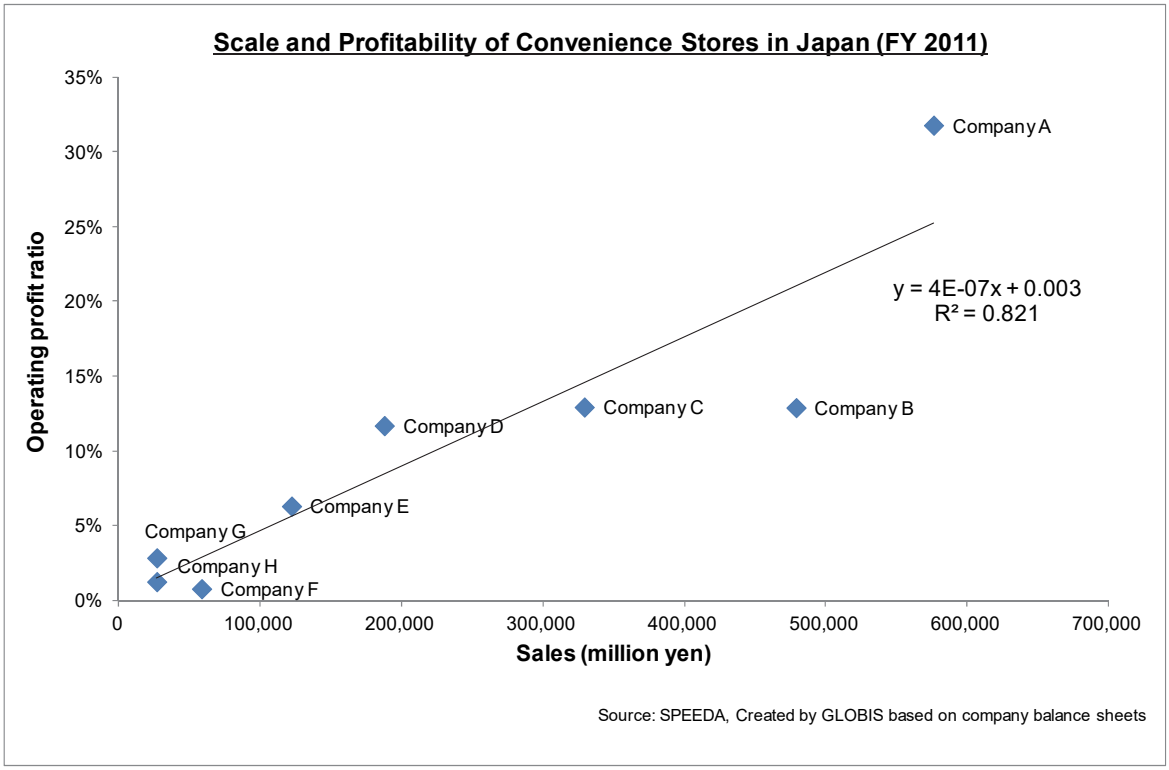


[2] The Power of Habit: Why We Do What We Do in Life and Business, Charles Duhigg

2) Perspectives on Analyzing the Output/Input Mechanism

Business is nothing less than the continuous pursuit and implementation of methods to realize the goals that have been set forth in a business.

Goal-method and causality are inseparable. Causality between goal and method is a requirement, and, if you understand causality, you will be able to think up effective methods for achieving the goal. For example, if you find causality (economies of scale) between company scale (better cost efficiency) and increased profitability, it is a good idea to expand the scale of the company to increase profitability. In fact, these economies of scale are a major factor when investigating corporate merger decisions. For example, the graph below shows the relationship between the scale of convenience stores in Japan and profitability, and it is clear that the larger the scale, the higher the profitability. In the convenience store industry, scale expansion is an important factor for strategic decision-making.



Prerequisites for causality

Are sales the only factor for the operating profit ratio? Or is this variable only one of many factors? How do we know if this causality can be proven to be true? This is why we have to identify the prerequisites or which conditions exist in causality. The following are three frequently used requirements.

1. In terms of time, the cause precedes the outcome
2. There is correlation (covariance)
3. The correlation cannot be explained by other variables (third factors)

First of all, some caution is necessary as correlation exists where there is causality, but it is not always true that causality exists because there is correlation. Caution is necessary because it is easy to overlook the third factors mentioned in point 3. For example, there is correlation between increased ice cream and beer sales in summertime, but there is no direct causality such as having a beer makes you want to eat ice cream. Summer heat, i.e., temperature, is the

shared factor (third factor). There is correlation in the sense that both sell well when it is hot and both sell less when it is cool, but there is no direct causality.

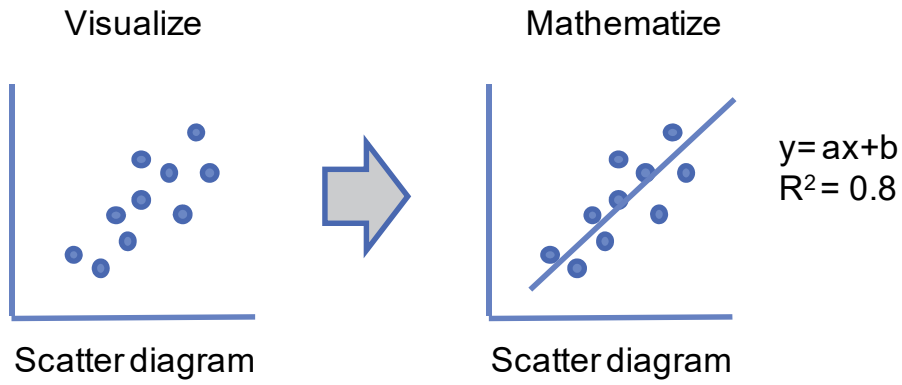
Presuming causality

As the discussion of prerequisites makes clear, whether or not correlation (covariance) exists is an important factor. In quantitative data analysis, it is important to confirm the absence or presence of correlation between variables. (Qualitative data can be represented as quantitative data through scatter diagrams and regression analysis by using the dummy variable approach, which will be discussed later.)

Correlation (covariance) analysis is commonly conducted through the following steps.

Visualization in a scatter diagram → (Concentration of the correlation coefficient)* → Quantification of the relationships based on regression analysis

* Analysis of the correlation efficient is included because the square of the correlation efficient becomes the coefficient of determination in regression analysis.



Unit 3-4: Approaches to Analysis

After going through the perspectives of quantitative data analysis, we need to choose the appropriate approaches.

Generally speaking, there are three approaches to analysis.

- Visual summaries (graphs)
- Numerical summaries (numbers)
- Formulaic summaries (formulas)

1) Visual summaries

You will not get much out of a large set of data by simply looking at the numbers. In reality, the eyes are the ultimate tools for analysis because they are excellent at recognizing patterns and other processes. A little manipulation of the data to create graphs makes it possible to extract meaningful information.

Here, we will leverage analytical perspectives by using graphs other than the simple pie charts and bar graphs that we all use on a daily basis. The course will focus on the following graphs, which are considered particularly useful for business.

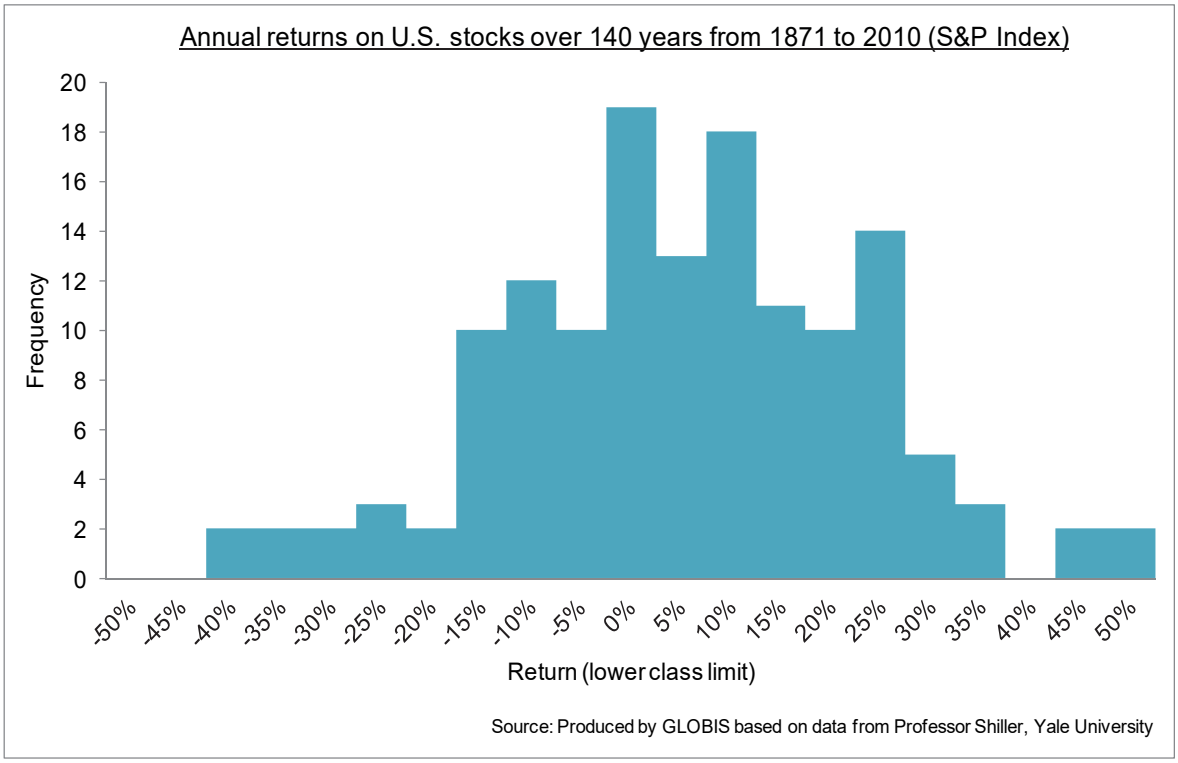
(A) Histograms
(B) Waterfall charts
(C) Pareto charts
(D) Time series graphs
(E) Scatter diagrams

(A) Histograms

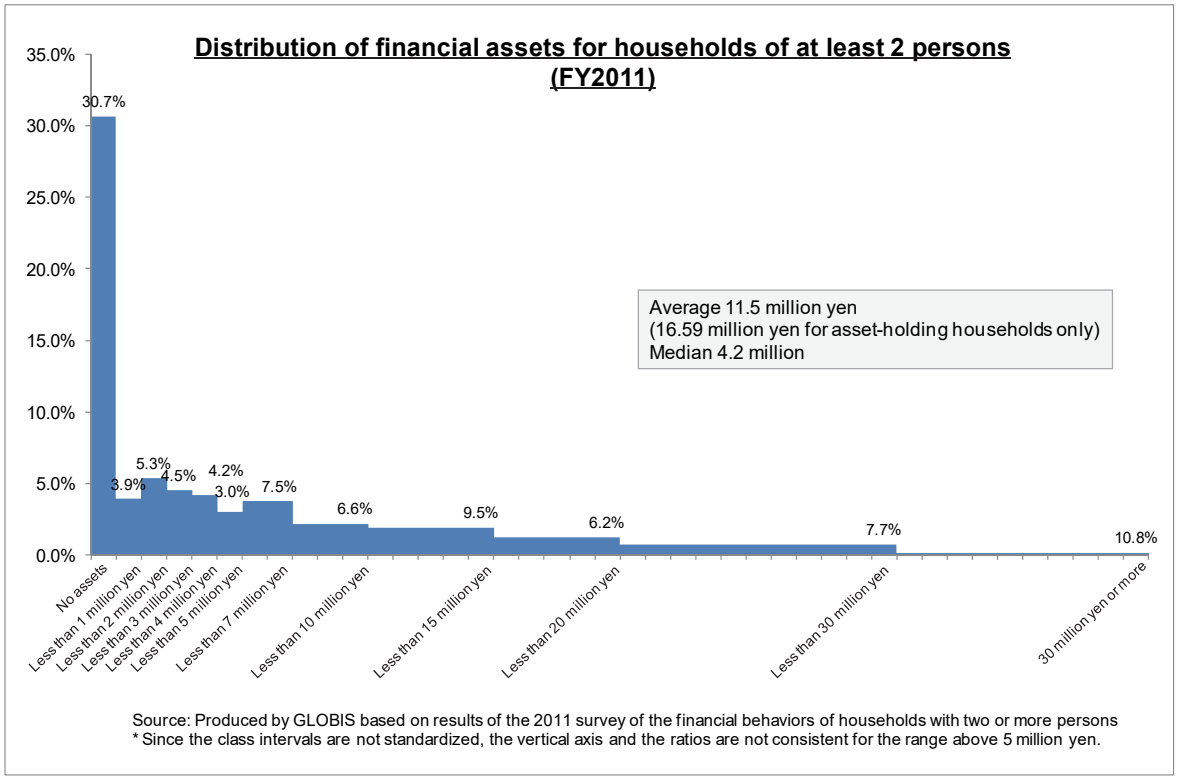
A histogram (frequency distribution chart) is a bar graph with the distribution of the variables plotted on the horizontal axis and the frequency (data instances) on the vertical axis. The area of the bars is proportional to the frequency. Histograms create a better visual understanding of the overall dispersion of the data.

If the data is distributed symmetrically in a bell shape, and if you have the mean value, which is the representative value, as well as the standard deviation, which indicates dispersion, it is possible to roughly understand the distribution. Subconsciously, we have a tendency to assume that distribution is symmetrical. However, in reality, as indicated by the distribution of assets illustrated in the graph below, or in the distribution of the scale of earthquakes, distribution in the universe is rarely symmetrical if ever, and usually skewed. Consequently, it is important to visually confirm the distribution.

The first graph shows the distribution of annual returns on U.S. stocks over a period of 140 years. We see that distribution is largely symmetrical around the average return of 8%. This histogram tells us that the -38.7% return after the collapse of Lehman Brothers in 2008 is positioned at the bottom end of the distribution, i.e., in terms of distribution, such events have occurred with a frequency of once or twice every 140 years.



On the other hand, in the histogram of the distribution of household financial assets below, the distribution is skewed to the left. Reading the histogram, we understand that the average value of household assets is 11.5 million yen, but the median (the value that separates the higher half of the data from the lower half) is 4.2 million yen. The main reason is that, proportionally, a few wealthy persons hold large financial assets, lifting the overall average value far above the median.



Depending on the width of the intervals, the shape of a histogram can change even when based on the same source data. Determining the number of intervals and their width is a critical point when creating histograms.

The answer to the question of what you want to find out from histograms is that by drawing histograms you want to find the "true distribution" of data based on the distribution in your sample data. You want to estimate the shape of the original distribution from limited numerical data extracted as a sample. Different sizes and numbers of intverals would create different outlooks of the data. To come back to the whole intention of wanting to find out the "true distribution" for data based on the distribution in your sample data, there are formulas for determining the number of intervals for an "appropriate" shape that is close to the true distribution. For example, Sturge's Formula[3] seeks the "logarithm to base 2 of n", $1 + \log_2 n$, for the optimal number of intervals, where n is the number of samples.

For example, suppose you want to find out the population distribution by age in Japan (the population pyramid). If you survey the whole population like the national census does, you will find out the exact population distribution, but suppose you want to get a degree of understanding of the population distribution for the whole of Japan (the population pyramid) based on age data for 100 randomly selected persons. If the interval widths are small as you draw the distribution based on increments of one year, the distribution will be uneven and the shape of the distribution will not be smooth because the sample is small. On the other hand, if you make the intervals widths extremely broad with increments of 20 years, peaks in the population, such as the baby boomer generation, will be submerged. The histogram will change drastically depending on how you handle the number and width of intervals. The term "appropriate" expresses what intervals should you choose to get a histogram with a smooth shape that is close to "true distribution" (think of it as the distribution for a very large number of samples).

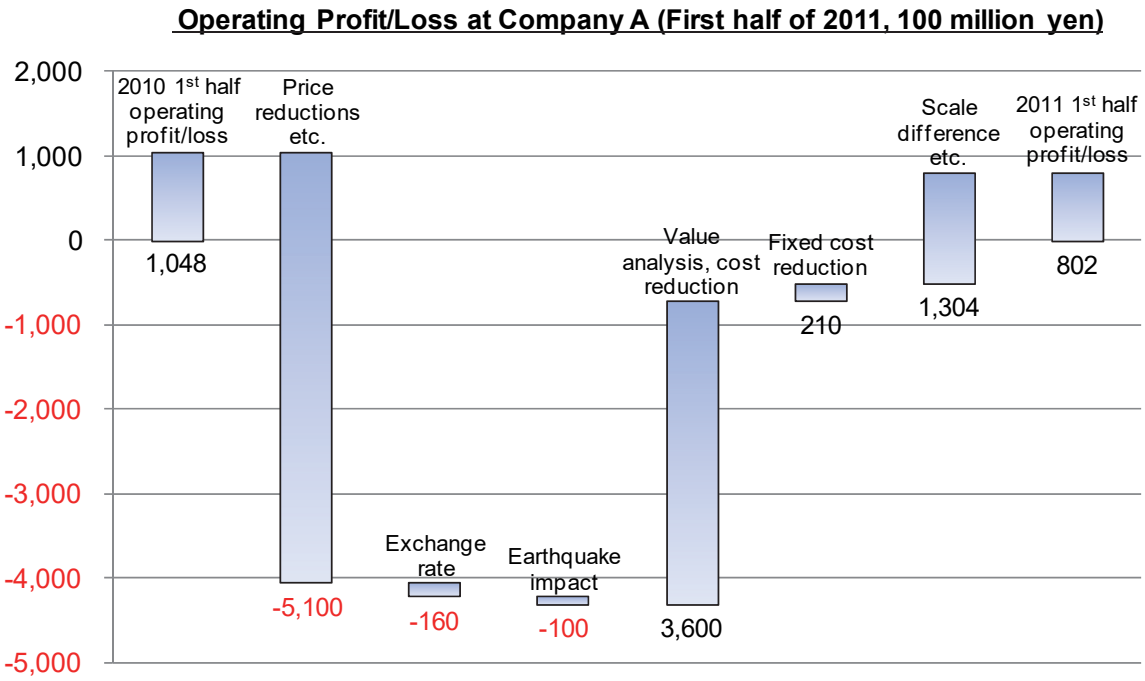
With Sturge's Formula, you get the number of intervals that creates a smooth shape close to a true distribution. In reality, as a guideline, it is safer to divide and change the number of intervals (generally in the range of 5-10), draw the graph and then check how unevenness in the distribution is expressed.

(B) Waterfall Charts

A waterfall chart is a graph that breaks down things that consist of several components in a steplike shape similar to a waterfall. By displaying the data in this way, it is possible to clearly show where the ratios are high, or which are the major factors among those that change over time. If it were only a matter of comparing compositions, you could also use a pie chart, but with a pie chart it is difficult to express changes over time, or elements with negative factors, which is where the waterfall chart gets its turn.

The waterfall chart is an extremely good graph for breaking down and visualizing changes and structures. Even though you can frequently see it in consulting reports, generally speaking, it is still a graph that you seldom come across.
This graph compares the financial results with the previous year at Company A, an electronics manufacturer. From the graph, we can understand that price reductions, the exchange rate and the impact of the Great East Japan Earthquake were factors that reduced profits with regards to operating income for the previous year, but to offset this, the company worked hard to reduce costs and improve their earnings structure.

[3] The number of bins $k = 1 + [\log_2 n]$, where [ ] is a ceiling function.



Source: Produced by GLOBIS based on the financial data of Company A

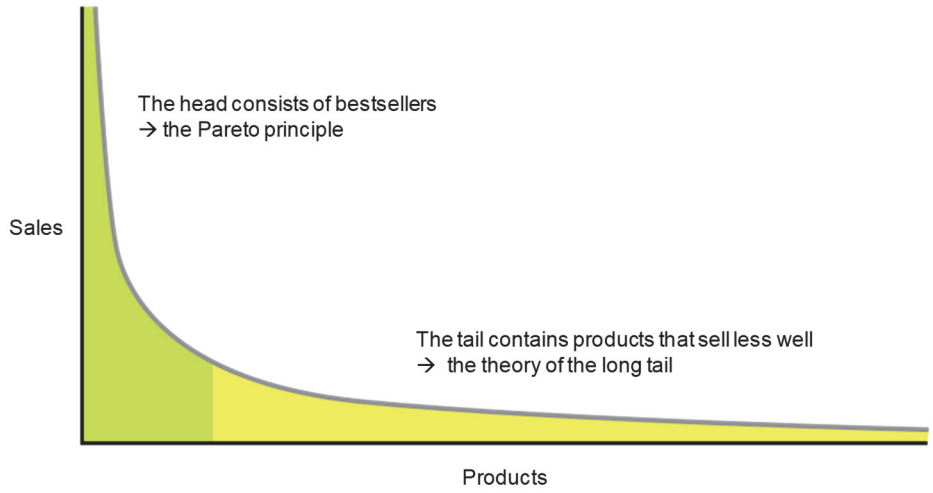
(C) Pareto Charts

The Pareto chart applies the Pareto principle, which states that there is bias in the world. Think of it as a histogram that sorts the factors in order of frequency (descending order). To make the bias clearer, we frequently use a broken line graph to note the cumulative distribution ratio in addition to the regular bar graph.

Pareto was a 19th century Italian economist who discovered that 80% of land in Italy was owned by 20% of the population. The Pareto Principle is also referred to as the 80/20 rule, which states that for many phenomena, 20% of the input produces roughly 80% of the results. To put it briefly, there is disproportion in the world. Since this means that focusing on the areas with the highest concentration is the most efficient way to get results, the Pareto Principle has long been used as a rule of thumb for many scenarios (you could write a whole book about the Principle alone).

In business as well, where resources are limited, it is impossible to obtain the maximum results by focusing on all factors. The most feasible approach is focus on the factors with the highest concentration (ideally on the top performing 20% of the total operation). So, this principle is used, for example, in quality management as well as to narrow down marketable products in the retail industry. In retail, POS analysis at convenience stores is a typical example of using the Pareto chart to focus on goods that sell well and to cut out items that move slowly, in order to maximize the use of limited shop space.

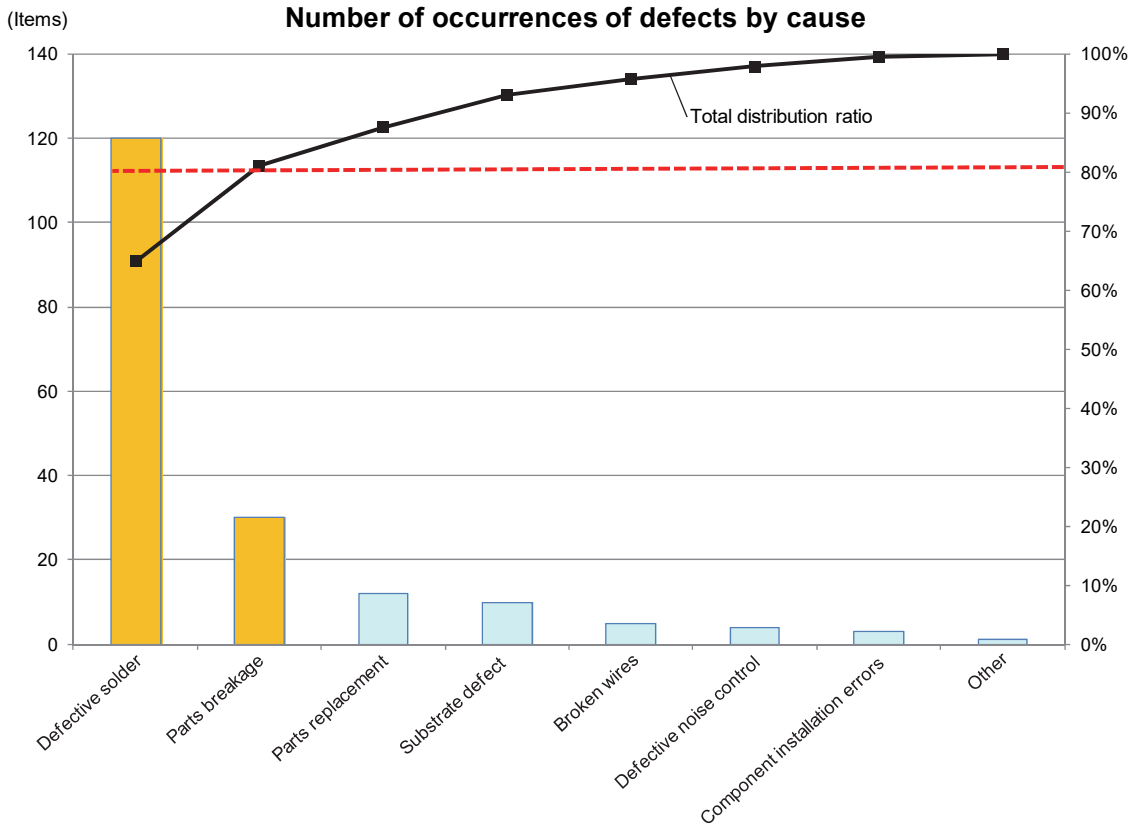
On the other hand, if there is no physical inventory restriction, such as with online vendors, less popular items (or, "the tail") may generate relatively high sales and profit compared to best sellers ("the head"). This is referred to as the theory of the long tail, or to use an old proverb, "Little and often fills the purse."



Source: Produced by GLOBIS based on Picture by Hay Kranen/ PD

Lining up the factors in descending order of quantity in a Pareto chart provides a list view of overall importance, the sequence of the items, and the cumulative values.

For example, in the following example taken from quality management, we can say that defective soldering and parts breakage account for 80% of defective products, hence improving them should be the priority.



(D) Time Series Graphs

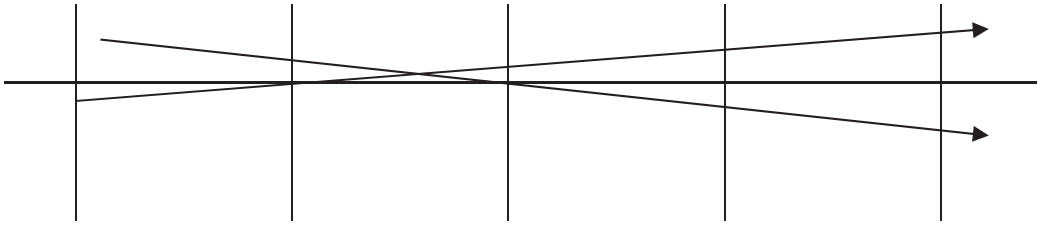
A time series graph measures time on the horizontal axis and data on the vertical axis to show changes over time. Normally, it is a vertical bar graph or a broken line graph, but when setting out multiple changes over time, the broken line graph is normally used instead of the vertical bar graph for ease of comparison.

There are several points of caution when setting out data in a time series. Generally speaking, time series data includes four components: trends, cycles, seasons and irregularities.

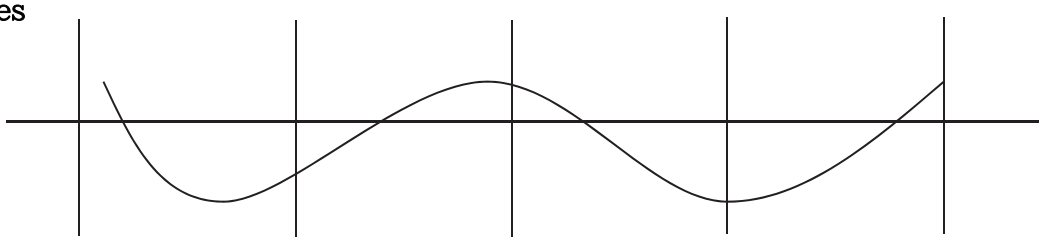
Trends: Long-term movement with an underlying direction (an upward or downward tendency) and rate of change.
Cycles: Regular, non-seasonal fluctuations over a certain time frame.
Seasons: Any regular fluctuating variation that associates with the seasonality of a period of less than one year.
Irregularities: Random fluctuations that cannot be explained by any of the three components above.

Since the implications of each of these variations are different (by different implications, we mean that the courses of action vary if you have to respond in some way), it is important to segregate them as much possible when interpreting the analysis.

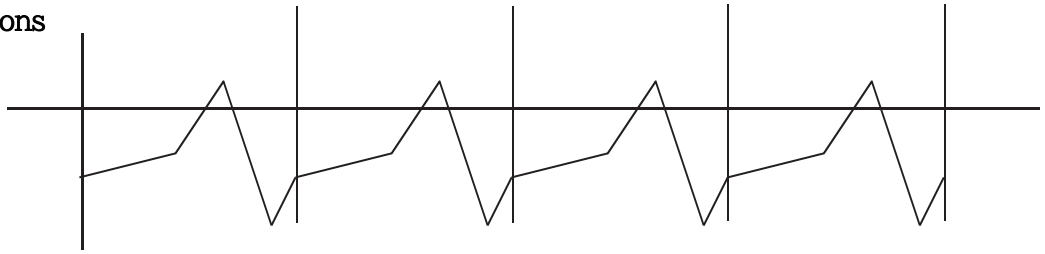
Trends



Cycles

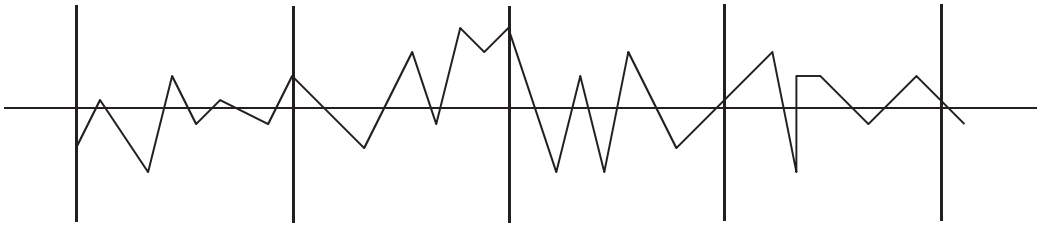


Seasons



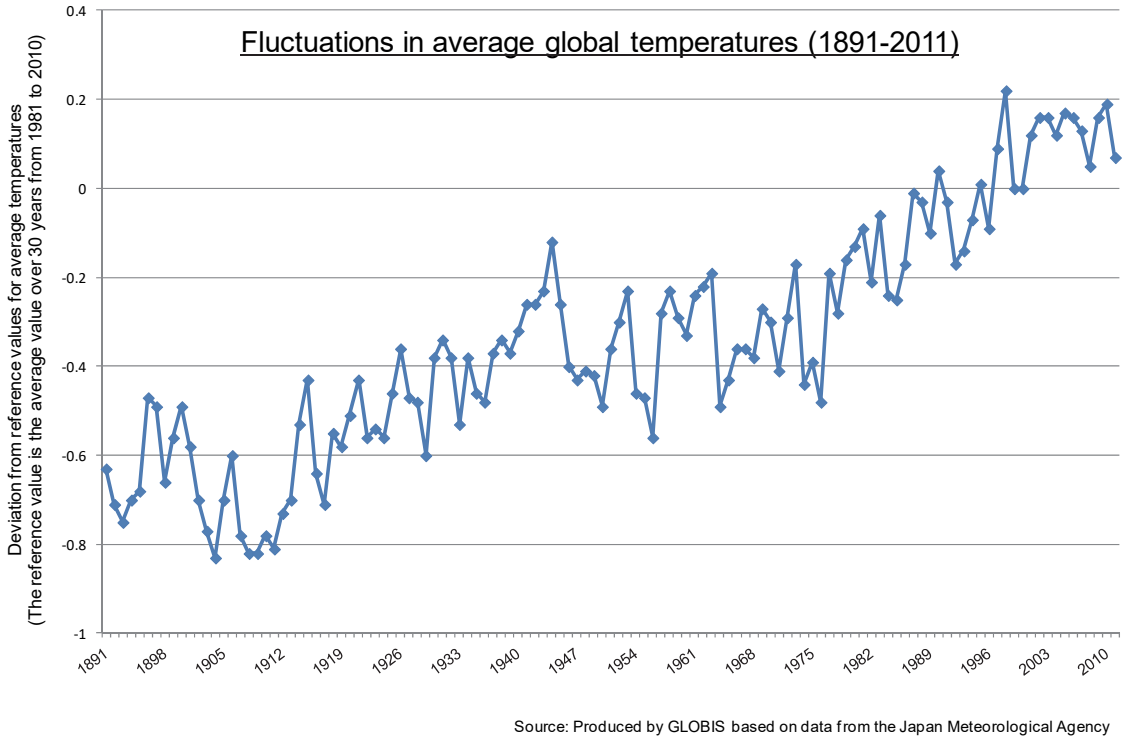
Irregularities also include phenomena with major impact, such as major systems changes (for example, large-scale deregulation) in a particular business, or ground-breaking technology development.

Irregularities



Determining trends from a time series is extremely effective when making forecasts about what will happen in the future. At such times, you first make forecasts by focusing on the trends and, if necessary, also take account of other variations such as seasonal changes.

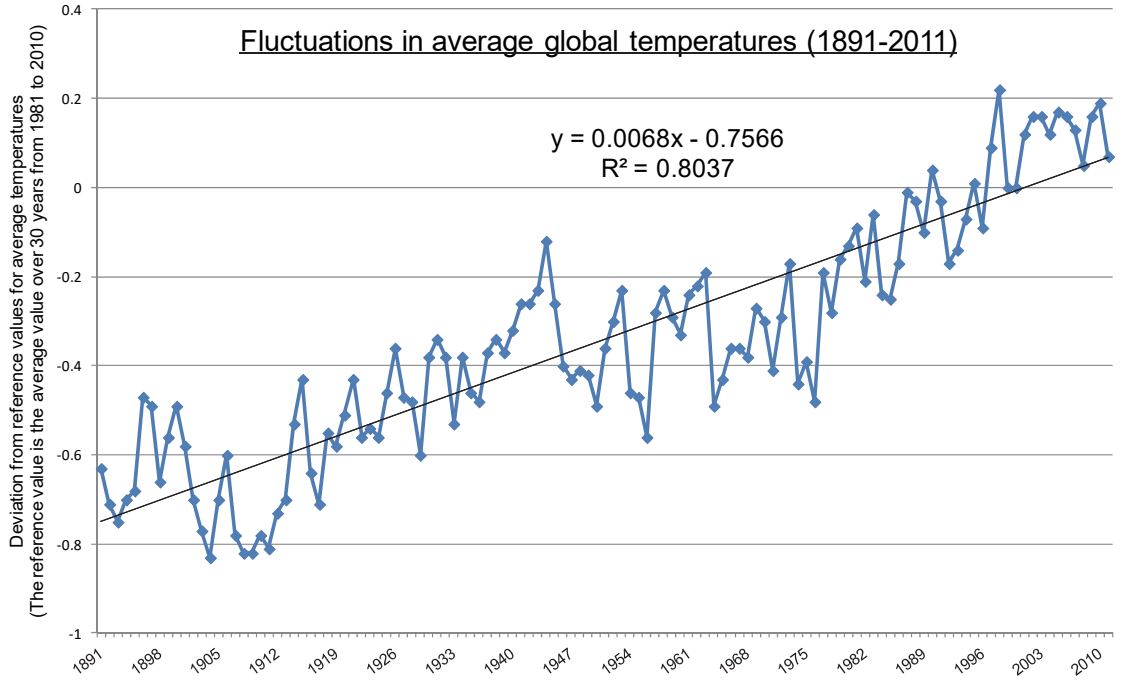
The trendline function in Excel (which simplifies trends) is a good way to extend a trend in a graph and use to get clear understanding of what happens next. For example, consider the questions, "Is our planet warming up?" and "Assuming global warming continues unchanged, how high will the temperature rise in 100 years?"



Source: Produced by GLOBIS based on data from the Japan Meteorological Agency

To start with, when you draw a time series graph, you can see for yourself that there is a rising trend. But how much of a rising trend is it? Right-click on the data in the graph and select >Add Trendline >Linear. At this time, make sure that the "Display Equation on chart" and "Display R-squared value on chart" are checked.

The result should be something like this graph.



Source: Produced by GLOBIS based on data from the Japan Meteorological Agency

From the incline of the formula shown in the graph, we understand that the trend is for temperatures to rise by 0.0068°C per year. If this trend continues, we can multiply the value by 100 to forecast an average temperature increase of about 0.68°C in 100 years of time. (This trendline is a simple linear regression analysis where the objective variable is the temperature and the explanatory variable is the year. Unit 3-4. 3) (A) will discuss simple linear regression analysis in more detail.)

(E) Scatter Diagrams

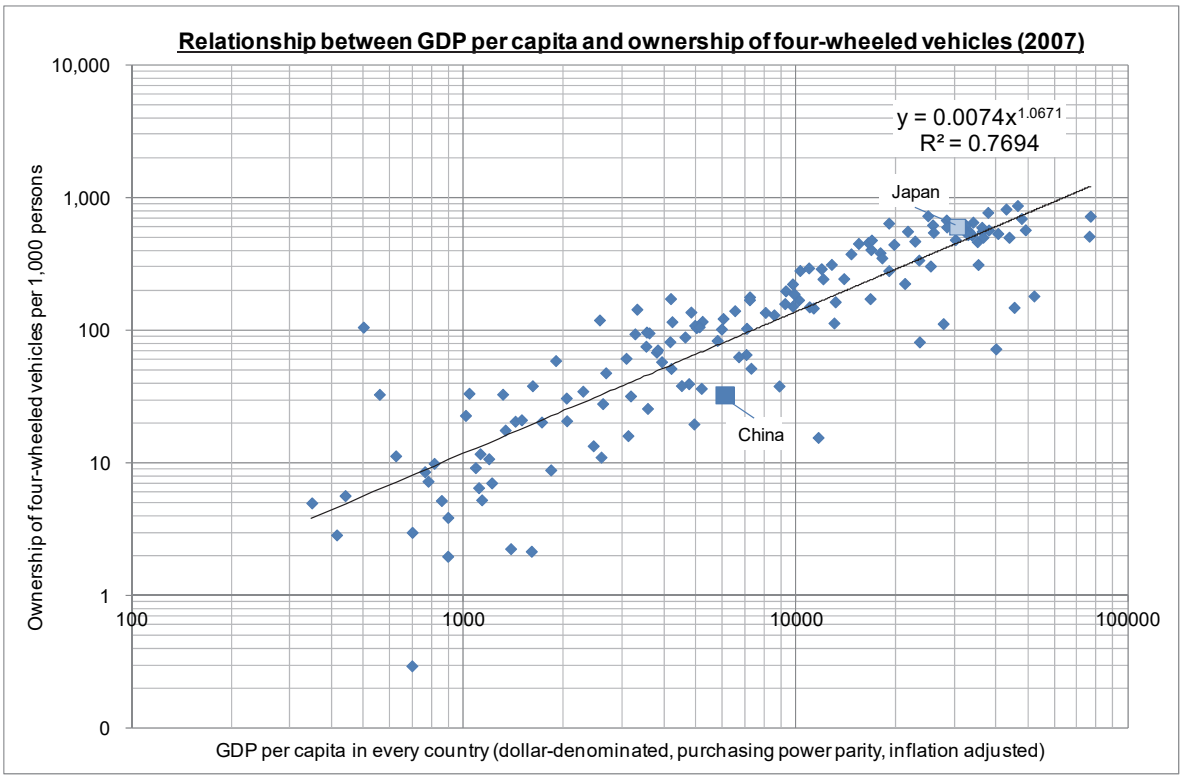
A scatter diagram is a graph for finding relationships between two variables by plotting the data against the variables on both the horizontal and vertical axes.

Of all the various graphs, this is one of the most important because of its impact and frequency of use, especially in science. If we probe both the economies of scale and the experience curves often alluded to the business, we will find that these relationships are interpretations of scatter diagrams. With the exception of the time series graph, all the graphs have so far basically dealt with a single variable, but the scatter diagram can be used to visualize the relationship between two variables.

For example, we can achieve the following objectives from a scatter diagram.

(1) If we understand the correlation between two variables from the trend, we can guess at the relationship between the two variables, including the causality (we can infer the output/input mechanism).

(2) We can get hints about data groupings from data aggregation.



Source: Produced by GLOBIS based on data from Gapminder and elsewhere

This scatter diagram shows the relationship between per capita GDP and ownership of automobiles in all countries. We understand that as people become affluent, there is a rapid spread in automobile ownership (note the graph has a logarithmic scale on both axes). For example, from this graph we understand how motorization in China will track future economic development.

When drawing scatter diagrams, and especially if we can conjecture causality, the causal series (input) is normally arranged on the x axis and the effects series (output) on the y axis. (Further, as in cross-tabulation, when making notations on the chart, causes are arranged as row variables and effects as column variables.)

2) Numerical summaries

This approach tries to understand complex matters by summarizing them in numbers in the simplest possible way. Broadly speaking, there are two important ways to summarize the data using numbers. By understanding the traits and values of both measurements, you will have a better idea of the overall state of the data.

(A) Where the center of the data is (representative value)
(B) How the data points are dispersed (level of dispersion)

(A) Where the center of the data is (representative value)

Simple average, weighted average, and geometric average (average annual growth rate)
The most frequently used representative value is the average value. Broadly speaking, there are two types of averages, the simple average and the weighted average.

The simple average is derived by adding up the numerical values of your samples and then dividing the sum by the number of samples. Weighted average, on the other hand, is derived by multiplying the numerical value of your samples by a weight, and then dividing the sum total with a weighted number.

For example, suppose that Company A has 30,000 employees who will get a 1,000 yen wage increase while the 5,000 employees of Company B will get a wage increase of 5,000 yen. If we calculate the average wage increase for Company A and Company B, the simple average is derived as follows:

$$\frac{(1{,}000 + 5{,}000)}{2} = 3{,}000 \text{ (yen)}$$

But, the weighted average approach, which is weighted by the number of employees, is derived as follows:

$$1{,}000 \times \frac{30{,}000}{30{,}000 + 5{,}000} + 5{,}000 \times \frac{5{,}000}{30{,}000 + 5{,}000} = 1{,}571 \text{ (yen)}$$

Normally, the simple average is used, but the weighted average may also be used when it is judged that each sample value will have a different impact on the average value. Some examples of weighted averages that often turn up in daily life or in business settings include the TOPIX (the Tokyo Stock Price Index: the stock prices of listed companies weighted against the number of shares issued), the consumer price index (the price of commodities weighted against consumption expenditure), and the concept of WACC in finance (the cost of debt and equity weighted against the volume of financing).

In addition to the simple and weighted average, the Compound Annual Growth Rate (CAGR) and average annual yield are commonly used averages in business settings. Unlike the averages mentioned above, they are not the sum of the growth rates for each year (yield) divided by the number of years, but are based on the geometric average, which is calculated in the following example:

[Example] Company C had sales of 10 billion yen 3 years ago. For the past three years, year-on-year growth rates at the company have been 20%, 90% and 10%. If the average annual growth rate is x, calculations show that x is 35.8%.

$$10 \text{ billion yen} \times (1 + x)^3 = 10 \text{ billion yen} \times 1.2 \times 1.9 \times 1.1$$

$$1 + x = (1.2 \times 1.9 \times 1.1)^{\frac{1}{3}}$$

$$x = 0.358$$

Even if you do not know the exact year-on-year growth rate for each year, you can calculate the average annual growth rate if you know what sales were three years ago (10 billion yen) and what they are for the current year (25.08 billion yen based on the abovementioned growth rates).

$$1 + x = (250.8/100)^{\frac{1}{3}}$$

$$x = 0.358$$

Median, mode

When distribution is in the shape of a bell curve around a central average, the average value is the number with the highest data concentration, which is very convincing for representing the data. But when there is bias in the distribution of data as in the case of the distribution of financial assets, the average value is not necessarily a convincing representative value. In such cases, the median and the mode better represent the overall picture in addition to the average value.

When you arrange the numerical values of the samples in order of magnitude, the median indicates the value that is at the midpoint of the number of samples (if you have an even number of samples, you take the average of the two numerical values at the midpoint).

For example, if you have 100 samples, the median is the average of the values of the 50^{th} and 51^{st} samples. The median is often used where it is expected that the majority of the samples will not be distributed around the simple average due to the characteristics of the group (if there is no bilateral symmetry around the simple average value when you draw the histogram). For example, in the case of financial assets per Japanese household, the simple average value is higher than the representative value for the group as a whole because of the impact of a few wealthy people. Consequently, we use the median when we talk about the financial assets of a typical Japanese household.

The mode refers to the numerical value with the highest frequency. It is used in cases where the data forms two or more peaks when you draw the histogram, or where outliers (exceptional values) influence the simple average.

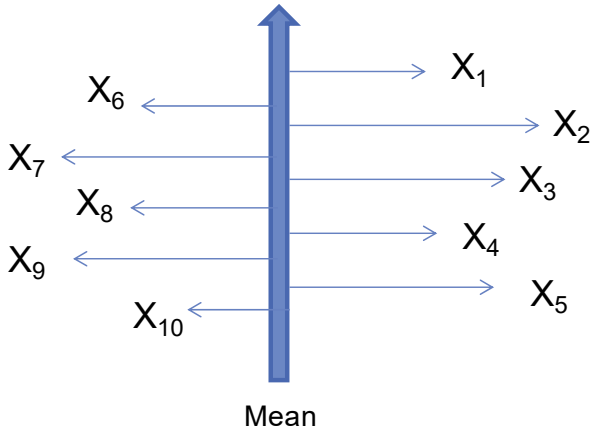
(B) How the data points are dispersed (level of dispersion)

Variance and standard deviation

Averages are very useful as representative values for large amounts of data, but they do not tell us anything about the overall distribution and dispersion of the data.

For example, suppose you have to pick a representative for the Olympics in swimming. But, unfortunately, none of the candidates have posted average times in the past year that are anywhere near the top three times in the world this season. If winning a medal is an absolute must, should we choose the athlete with slightly higher average times, but little dispersion in competition data, or the athlete with inferior average times, but more dispersion in the data? In this case, it might be better to gamble everything on the athlete with more dispersion in competition times than the athlete with better average times.

The pitch of ocean waves describes the dispersion in the level of the sea. Fishermen that make their living on boats, or surfers, are not interested in the tide level, which is the average height of the sea level, but in the waves, which represent dispersion. If you are an offshore fisherman, you hope for a sea that is not stormy (little dispersion in sea levels), while surfers hope for large waves (great dispersion in sea levels).
We can use variance and standard deviation to look at the state of dispersion in this data.



In the graph above, the data is dispersed around the mean. Suppose we want to know how each data point is dispersed around the mean. For obvious reasons, some data points are distributed further from the average, and other data points are closer. Consequently, when we seek the difference between the data points and the mean, we get both positive and negative values. Even if we only try to take the average of the differences between the data points and the mean, in other words deviations, the positives and negatives offset each other and we end up with zero. Therefore, we would take the average of squared deviations from the mean, which is called variance (σ^2).

$$\sigma^2 = \frac{\{(X_1 - \bar{X})^2 + (X_2 - \bar{X})^2 + \cdots (X_n - \bar{X})^2\}}{n}$$

* $\bar{X}$ is the mean value and n is the number of samples. When dealing with samples rather than populations, the denominator is not n, but n-1.

Meanwhile, standard deviation (SD or σ) is derived by taking the (positive) square root of the variance (in short, undo what we just squared). Think of standard deviation as denoting average dispersion and separation of a set of data from its mean. Since we have squared the

original number and then taken its square root, the unit is as easy to understand as the original figure, so standard deviation is used more often than variance as an estimate of dispersion.

$$\sigma = \sqrt{\frac{\{(X_1 - \bar{X})^2 + (X_2 - \bar{X})^2 + \cdots (X_n - \bar{X})^2\}}{n}}$$

* $\bar{X}$ is the average value and n is the number of samples. When dealing with samples rather than populations, the denominator is not n, but n-1.

Compared to the mean, standard deviation may seem unfamiliar. However, as mentioned previously, in business, it is a value with implications that are far more important than the average. For example, minimizing product quality variance, or standard deviation, and measuring homogeneity were important features of the outstanding quality control in operations that characterized Japanese business in post-World War II Japan. Standard deviation also plays an important role for risk concepts in finance.

The following is a summary of the advantages of standard deviation:

1) Understand position in the dispersion
→ (When normal distribution is assumed), determine the presence of any unique examples
2) Able to compare different distributions
→ Allows standardization according to distance from the mean. Deviation value on test scores is a typical example.

1) Position in the dispersion (Is the data common or unique?)
We know that for a range of natural and social phenomena, distribution that follows the bell curve of normal distribution is an extremely powerful distribution. In many cases, the normal distribution is a reasonable approximation for a prior probability distribution for business decision purposes, and is used in many applications. It should not be assumed that every process has a normal distribution.

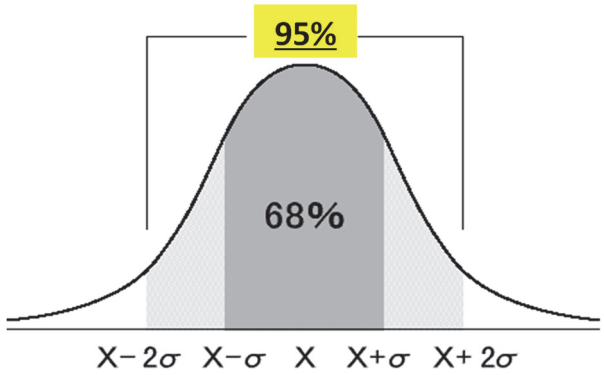
Despite the difference in normal curves, they all share an important property that allows us to treat them in a common way. The mean, X has a universal relation with the standard deviation, σ in all normal distribution which can be mathematically expressed as the following:

(1) The range $X - \sigma \leqq X \leqq X + \sigma$ comprises 68.3% of overall distribution
(2) The range $X - 2\sigma \leqq X \leqq X + 2\sigma$ comprises 95.4% of overall distribution
(3) The range $X - 3\sigma \leqq X \leqq X + 3\sigma$ comprises 99.7% of overall distribution

That is to say, since nearly two thirds of all data points fall within plus or minus one sigma of the average, we understand that data that deviates by plus or minus one sigma from the mean is fairly common. On the other hand, since 95% of data falls within plus or minus 2 sigmas of the mean, we understand that any data that deviates by more than plus or minus 2 sigmas from the mean is rare.

As we will discuss later, people have a tendency to see a probability of less than 10% (for example, 5%) as something that is unlikely to occur. Consequently, there are many cases where the question of whether data falls within the first two standard deviations (95% probability) is used to determine whether or not something unlikely has occurred.

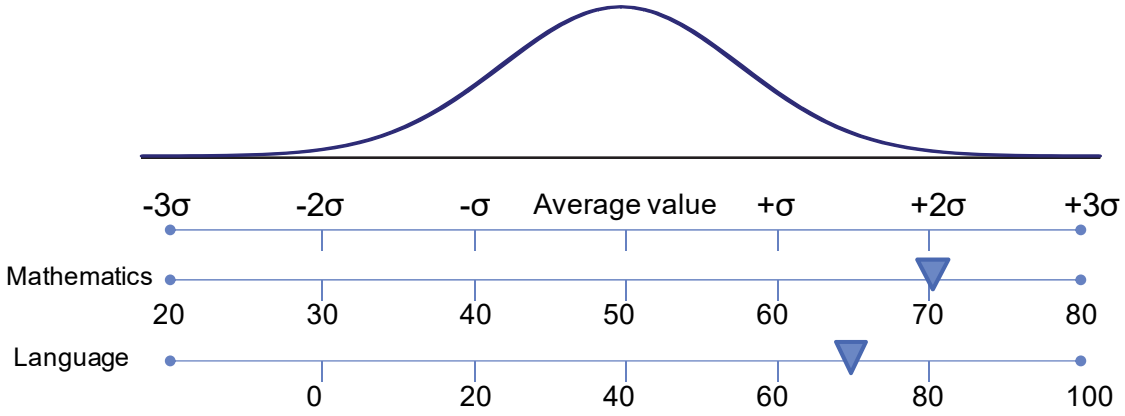
Remember that whether or not data is within two standard deviations of the average is referred to as the 2 SD rule.



2) Compare distributions (Which are further away from the mean?)
For example, suppose you scored 70 on both tests in mathematics and linguistics. Supposing the mean score for the math test was 50 and 40 for the linguistics. Which test did you perform better in? For mathematics, your deviation was 20 points, but for the linguistics test, your deviation was 30 points, so did you do better on the linguistic test? When comparing different distribution intervals, you need something for comparison. Standard deviation, which measures dispersion from the mean, is used to draw comparisons. Assuming we know that the standard deviation for the math test is 10 points and the standard deviation for the linguistics test is 20 points, your scores are:

Math:	+ 2 standard deviations from the average
Language:	+1.5 standard deviations from the average

So, from the location of the distribution, we understand that the mathematics score ranks higher.



Standard Score, which is called Hensachi in Japanese, is widely used in Japanese examinations where the mean is 50 points and the standard deviation is 10 points. In the case of the above example, the standard score for mathematics is 70 points and the standard score for linguistics is 65 points. Standard score itself has been blamed for the intensification of the exam wars, but originally, it was introduced as a device for comparing different distributions.

Is Japan a homogenous society? (Using standard deviation to look at homogeneity in Japan)

It has often been said that Japan is a homogenous society. A recipient of the Order of Culture of Japan, Chie Nakane, suggested the same in her book, “Human relations in a vertical society”

(in Japanese, Tateshakai no ningen kankei). Therefore, many Japanese think of themselves as ethnically homogenous.

So, let's have a look at the data to see if Japanese people are, in fact, homogenous. To start with, the target for comparison is important here. To find out whether Japan is homogenous or not, we compare with other countries, so let's draw some comparisons with dispersion in other countries to find out if dispersion in the values of Japanese people is relatively high or low compared to other countries. If Japanese are ethnically homogenous (compared to other countries), the standard deviation in Japanese people's values should be extremely small compared to standard deviation elsewhere.

Fortunately, it is possible to compare data because we have past cross-sectional data surveys of the values of people in all countries, such as the International Social Survey Program and the World Value Surveys. These two surveys examine a wide range of values including the environment, family/gender, state, government, occupation, religion, etc.

According to research[4] by Mabuchi, among the 223 survey items that concern values, there are only three items (1.3%) where Japanese are extremely homogenous compared to other nations, i.e., where the standard deviation is small. On the other hand, there are 21 items (9.4%) where the standard deviation is extremely high compared to other nations. The remaining 199 items (89.2%) had about the same degree of standard deviation as other countries.

Items where the standard deviation of Japanese responses was extremely small compared to other countries.

Survey	Question
ISSP	Would you like more or less time for housework?
	Would you like more or less time with your family?
WVS	Social norms: Is tax evasion right (acceptable)?

Some items where standard deviations for Japanese responses were extremely high compared to other countries.

Survey	Examples of questions
ISSP	Children suffer psychologically if mothers work outside the home before the start of schooling
	Should both men and women earn money toward the household finances?
	Is life without children empty?
	Should the government spend more money on insurance and healthcare than they do now?
	Do you think ordinary citizens also have an influence on politics?
	Are you proud of the work you do in your current workplace?

As far as standard deviation is concerned, we cannot say that the values of the Japanese people in the survey were particularly homogenous compared to other nations. Instead, the data tells us that dispersion is high for values that involve family/gender, government and occupation.

[4] *Nijibunseki niyoru nihonjin doushitsuron no kensho* [Homogeneous Japanese?: A Critical Assessment with Secondary Analysis], 17(1), 3-22, 2002.

3) Formulaic summaries

Formulaic summaries are powerful yet simple ways of expressing the relationships between inputs and outputs, and are therefore very common in analysis.

(A) The Inductive Approach

Inferring what kind of input factors determine the output, their mechanism and causality is a given for problem-solving. It is also very important in the field of business and indispensable when considering choices of efficient methods to achieve targets when formulating strategy.

The most common method for data analysis is by using the visualization of the scatter diagrams (previously discussed), and regression analysis to formulate the mechanisms. By expressing the relationship between output and input in a numerical formula, it is not only possible to infer causality, but also, for example, to quantitatively predict the output based on the inputs, such as sales prognoses for new stores.

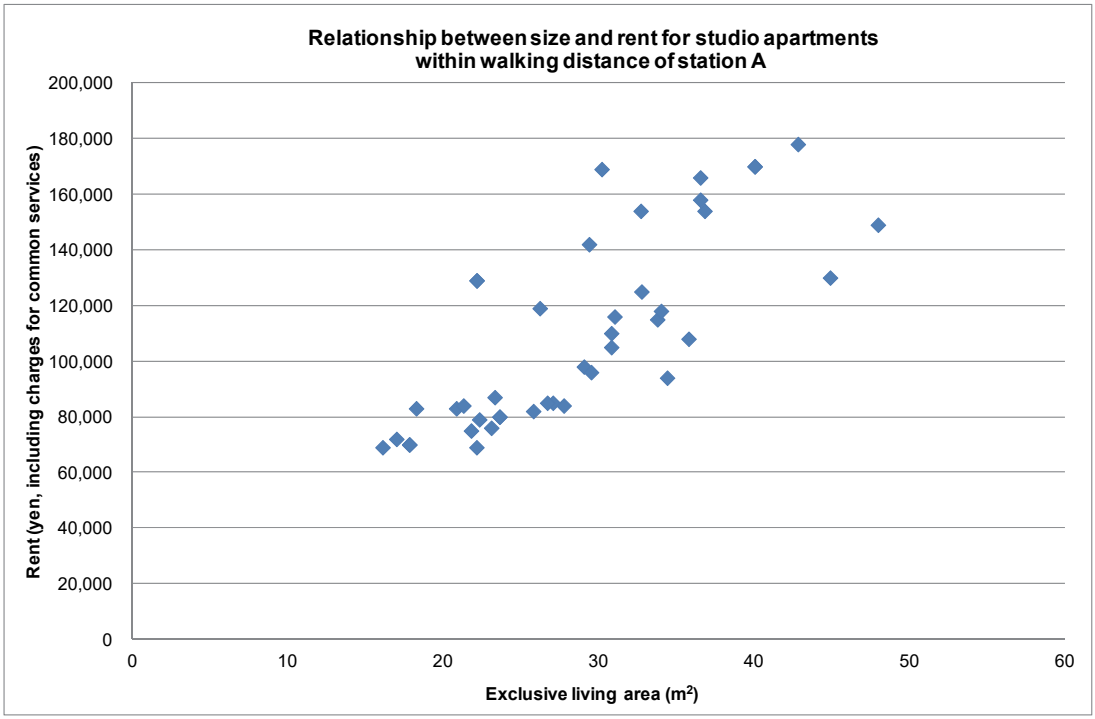
As previously mentioned, the economies of scale and experience curves often referred to in business also visualize the relationship between output (cost) and input (scale and cumulative production volume) in scatter diagrams, and use regression analysis to estimate the relationships expressed in numerical formulas. Most of the frequently used rules in business are concerned with the relationship between output and input, and can be visualized in a simple way with scatter diagrams or formulated with regression analysis.

Here, we will learn about regression analysis (simple linear regression, multiple regression) and the use of scatter diagrams for visualization and correlation.

Correlation and simple linear regression analysis

Now, suppose you want to buy a studio apartment close to station A in the city center and that you are planning to use the rental income to cover living expenses. The studio apartment is 25 m^2. So, first you have to think about the mechanisms that determine the rents for studio apartments.

Assuming that rent is determined by the size of the studio apartment; we have accessed real estate websites to collect rental data for 42 studio apartments in the neighborhood around station A. Let us understand the data further by visualizing it. The following graph shows the data in a scatter diagram.



The graph function in Excel can be used to easily create scatter diagrams.

When creating scatter diagrams, the cause (the input variable) is plotted to the x axis and the effect (the output variable) is plotted to the y axis. From this graph, we understand that the larger the apartment, the higher the rent, or to rephrase, there is "positive correlation" (the line rises to the right) between the rent and the apartment size.

Correlation refers to a state where there is some kind of regularity or linkage between two variables. For example, if beer sales increase when temperatures rise, and fall when temperatures drop, we can say that there is correlation between temperature and beer sales. Correlations can be positive/negative and strong/weak, which is expressed in a numerical value referred to as the correlation coefficient.

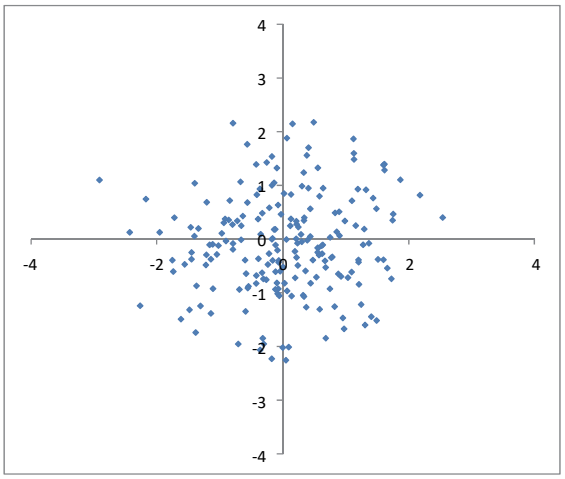
A +1 correlation coefficient means a perfect increasing linear relation while a -1 means a perfect decreasing linear relation. If the value is close to 0, there is almost no relationship between the variables. So, the closer the correlation coefficient is to either -1 or +1, the stronger the correlation between the variables. In business, the typical acceptable value for a strong correlation is 0.7 or higher, depending on the individual analytical requirements. The squared value of the correlation coefficient is the coefficient of determination, which will be explained when we discuss simple linear regression analysis.

[Interpreting the numerical values of the correlation coefficient (absolute values)]

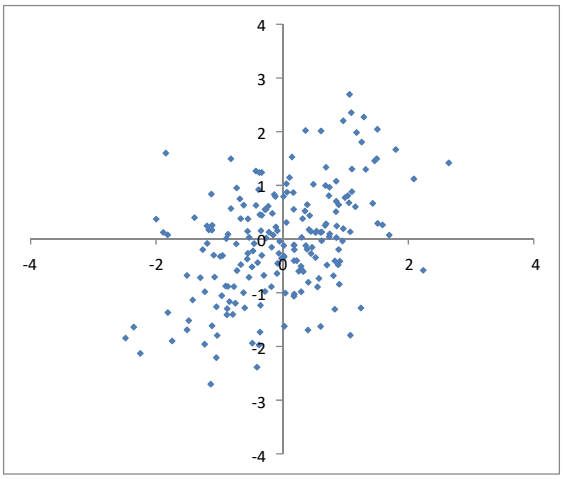
$-0.2<0<0.2$:	Negligible correlation
$-0.2<-0.4$ and $0.2<0.4$:	Weak correlation
$-0.4<-0.7$ and $0.4<0.7$:	Moderate correlation
$-0.7<1.0$ and $0.7<1.0$:	Strong correlation

Let's look at some examples of what each correlation coefficient actually looks like in a scatter diagram.

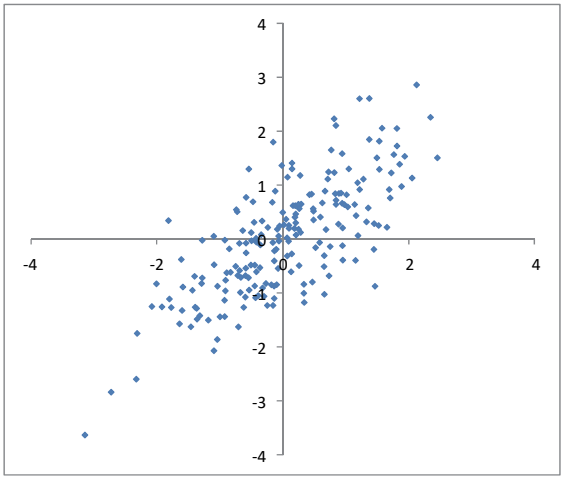
R=0



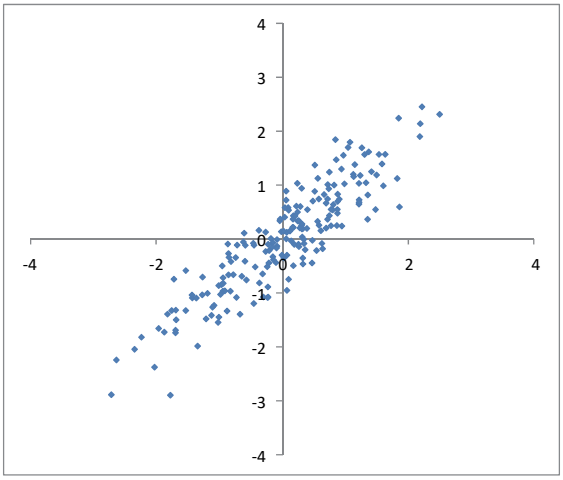
R=0.5



R=0.7



R=0.9



Now, let's try to use the data analysis tool for correlation in Excel to find the correlation coefficient for the two variables for studio apartments.

	Exclusive living area (m^2)	Rent + charges for common services (yen)
Exclusive living area (m^2)	1	
Rent + charges for common services (yen)	0.778038041	1

We find that the correlation coefficient for rent and living area is 0.78, which is a strong correlation.

The correlation coefficient plays an important role in similar situations. When you buy something from an online store, you often see that many sites have a recommendation function that says something like "We have recommendations for you." This is where the correlation coefficient is used. By calculating the correlation efficient between your purchase or browsing histories and those of other customers, the system compares customers with a high correlation coefficient, i.e., customers with a similar purchase and browsing history to yours, and presents the difference as recommendations.

Simple linear regression analysis

The correlation coefficient indicates the strength of the relationship between two variables, but it does not say anything about the living area factor in terms of monetary impact. Regression analysis formulates and analyzes this relationship. Where business is involved, we have to consider multiple factors when we try to explain phenomena that are generally complex. With regression analysis, we could explain phenomena by using numerical formulas that include several factors like the one below:

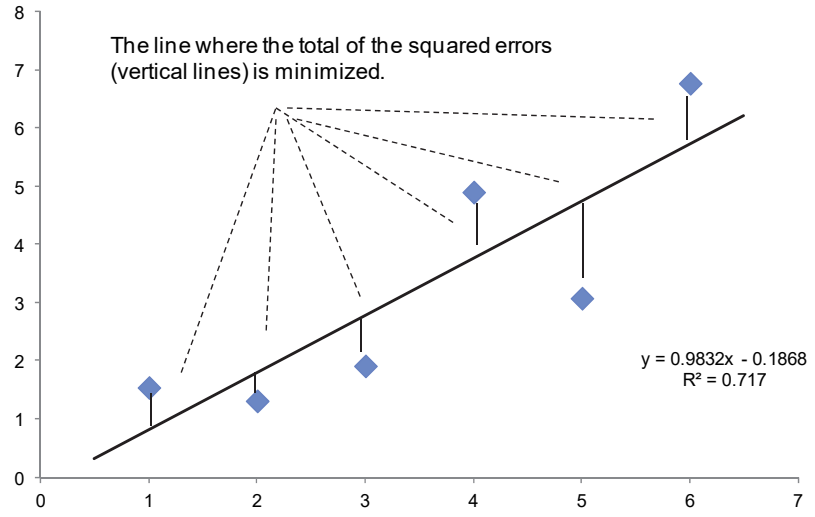
$$y = a_1x_1 + a_2x_2 + a_3x_3 + \cdots + a_kx_k + b$$

In this example, we have used a linear expression, but there are also formulas that use logarithms and exponents, or polynomials where x is raised to a power, such as x^2, x^3 and so on. However, in business, the linear expression is often used for its simplicity. Since the y on the left side changes with changes in the value of x on the right side of the equation, it is referred to as the objective variable, or the dependent variable, while x is the explanatory variable, or the independent variable. In addition, b is a constant term and a_k is the regression coefficient.

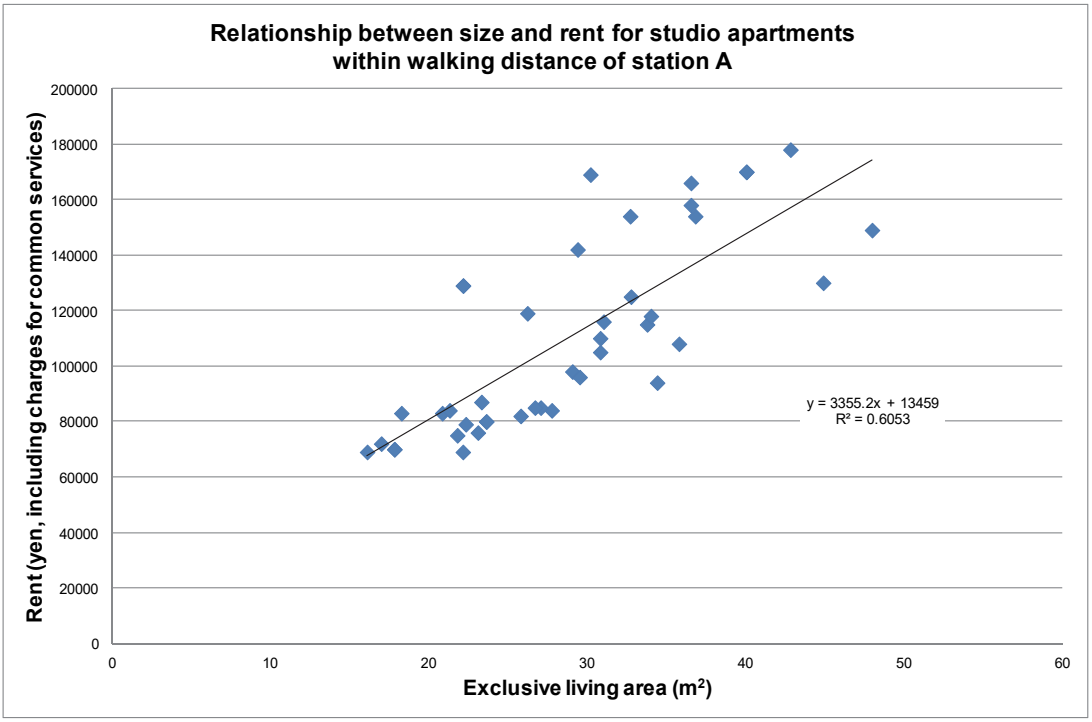
A simple linear regression analysis has one explanatory variable while a multiple linear regression analysis has several. Consequently, simple linear regression models are expressed with the following formula:

$$y = a_1x_1 + b$$

Visually, simple line regression analysis corresponds to drawing a straight line with the best fit through the data in the scatter diagram. Of course, you can try to manually draw a straight line through the graph, but the line will not be accurate and it will not be the same when you draw it the second time. In simple regression analysis, the best fit means to objectively draw a straight line so that the distance from each data point to the line is minimized. Specifically, as the graph below shows, it is the straight line that minimizes the sum of the squared lengths of the lines from each data point to the regression line. As much as we would like to minimize the sum total of the distances, we seek the regression line that minimizes the sum of the squares because of the ease mathematically.



Now, let's try to calculate the relationship between rent and living area with the regression tool in Excel. The results of the analysis and the graph are shown below.



SUMMARY OUTPUT

Regression Statistics	
Multiple correlation R	0.7780
Multiple determination R-squared	0.6053
Adjusted R-squared	0.5955
Standard error	21,730.3594
Number of observations	42

ANOVA

	df	SS	MS	F	Significance F
Regression	1.0	28,971,826,447.5	28,971,826,447.5	61.4	0.0
Residual error	40.0	18,888,340,792.9	472,208,519.8		
Total	41.0	47,860,167,240.5			

	Coefficients	Standard Error	t Stat	P-value	Lower 95%	Upper 95%	Lower 95.0%	Upper 95.0%
Intercept	13,458.7	12,758.5	1.1	0.2978	-12,327.1	39,244.5	-12,327.1	39,244.5
Exclusive living area (m2)	3,355.2	428.4	7.8	0.0000	2,489.5	4,221.0	2,489.5	4,221.0

At first glance, the results seem complicated, but there are three points you should look at.

- Intercept and coefficients
  The equation for best fit is shown in the lowest part of the chart. The intercept at the bottom is the constant term in the linear expression, and the exclusive living area is the coefficient for living area.

  As a result, the regression line is:

  Rent = 3,355 yen/m^2 × Living area (m^2) + 13,458 yen

  Based on this equation, we understand that the rent for the 25 m^2 studio apartment is 97,340 yen.

- Multiple correlation R
  This refers to the correlation coefficient previously mentioned. In order to interpret its significance, square the value to obtain R-Squared which will be explained below.

- R-Squared
  Generally, this is referred to as the coefficient of determination. If we read the coefficient of determination as a percentage, the value indicates the degree in which the explanatory variable explains the dispersion in the objective variable (how well it fits the regression), or in other words the explanatory power of the explanatory variable. We know that by squaring the correlation coefficient, we get a value from 0 to 1 (0≦R2≦1). Since the coefficient of determination in this example is 0.60, we say that 60% of the fluctuation in rent can be explained by living area.

Multiple regression analysis

In the example of simple linear regression analysis of rents, the coefficient of determination is 0.60 and from this we understand that 60% of the rent can be explained by the size of the living area. To further improve the explanatory power, let's consider the volume of information, specifically, by increasing the number of explanatory variables.

In multiple regression analysis, we use two or more explanatory variables to explain the phenomenon we want to clarify.

So, let's add the two explanatory variables of walking distance from the station (minutes) and how long since construction (years) as other factors than size that have an effect on the rent.

Using the Excel regression analysis tool produces the following results.

SUMMARY OUTPUT

Regression Statistics	
Multiple correlation R	0.9597
Multiple determination R-squared	0.9211
Adjusted R-squared	0.9149
Standard error	9,969.0133
Number of observations	42

ANOVA

	df	SS	MS	F	Significance F
Regression	3.0	44,083,680,660.4	14,694,560,220.1	147.9	0.0
Residual error	38.0	3,776,486,580.0	99,381,225.8		
Total	41.0	47,860,167,240.5			

	Coefficients	Standard Error	t Stat	P-value	Lower 95%	Upper 95%	Lower 95.0%	Upper 95.0%
Intercept	112,084.9	9,925.8	11.3	0.0000	91,991.1	132,178.6	91,991.1	132,178.6
Walking (min)	-2,988.3	526.2	-5.7	0.0000	-4,053.5	-1,923.1	-4,053.5	-1,923.1
Exclusive living area (m2)	1,847.2	233.7	7.9	0.0000	1,374.1	2,320.3	1,374.1	2,320.3
Age (years)	-1,639.3	149.5	-11.0	0.0000	-1,942.0	-1,336.6	-1,942.0	-1,336.6

Similar to the simple linear regression analysis, the results seem complicated at a glance, but there are four points you should look at.

- Intercept and coefficients
  The equation for best fit is shown in the lowest part of the chart. The intercept at the bottom is the constant term in the linear expression, and the coefficients for each of the variables are also noted.

  Based on this, the regression line is:

  Rent = -2,988 yen/minute × Walking (min) + 1,847 yen/m^2 × Living area (m^2) - 1,639 yen/year × Age (years) +112,085 yen

  The apartment you own is a 25 m^2 studio apartment at a 5-minute walk from station A and it was built one year ago, so, based on this equation, which takes distance to the station and years since construction into account, we understand that the rent is 141,684 yen. From this equation, we understand that the rent decreases by 2,988 yen for every additional minute of walking time, and by 1,639 yen for every year since it was built.

- R-Squared
  Similar to simple line regression, this is generally referred to as the coefficient of determination. It indicates the accuracy and explanatory power of the whole regression equation. If we read the coefficient of determination as a percentage, it indicates the ratio to which the explanatory variable explains dispersion in the objective variable (how well it fits the regression), or the explanatory power as it were. Think of it as roughly the degree (%) to which the explanatory variable explains the objective variable.

  As we already know from squaring the correlation coefficient, we get a value from 0 to 1 ($0 \leq R2 \leq 1$). Since the coefficient of determination is 0.92 in this example, we understand that 92% of the fluctuation in rent can be explained by living area, walking distance from the station, and years since construction. Compared to simple line regression, the explanatory power has significantly increased.

- Adjusted R-squared
  Generally, this is referred to as the coefficient of determination adjusted for degrees of freedom. The more explanatory variables are added, the more the coefficient of determination increases. So, when there are several regression equation candidates, we normally use the rise and fall in the adjusted R-square (the coefficient of determination adjusted for degrees of freedom), which compensates for the number of explanatory variables, for the multiple linear regression analysis and select the high regression equation. In case of multiple regression analysis, this value, not the coefficient of determination, is used to select the regression equation.

- P-value
  The P value is calculated for every variable and tells us whether the variable is statistically significant. This value is also called the risk ratio. Ideally, we are looking for a P-value as low as possible so there is less risk involved of misinterpreting the analysis. If the P-value is too high, we have to reject this variable as one of the explanatory variables. Generally, if the P value is higher than the significance level (a predetermined criterion for probability, often 10%) when expressed as a percentage (multiplied by 100), it is deemed high and best not used for the explanatory variable.

  In statistic testing, we build a (contradictory) hypothesis that the coefficient for the explanatory variable actually was null (= the explanatory variable has no significance), but since this causes inconsistency, we use proof by contradiction, which says that the

original coefficient was not null, but significant. Specifically, based on the (contradictory, actually null) hypothesis, the P value indicates observed probability for the current data. Generally, in business, there is hardly any possibility that the coefficient is zero if the significance levels (risk ratio) are 10% or below (0.1 or below), i.e., that the explanatory variable has significance. However, there is no statistical proof for the 10% criterion for risk ratio. It is simply customary to use it as the criterion for probability that is unlikely to occur.

People seem to feel that an event with a probability below 10% is unlikely to occur. For example, suppose someone is tossing a coin in front of you and asks you to pick heads or tails. Suppose that for some reason, heads keep coming up. How many times in a row can heads come up before you become suspicious? From experience in the class, most people get suspicious when heads come up 4 or 5 times in a row. Since the probability that heads will come up 4 times in a row is 0.5 multiplied by 4, or 6.25%, and the probability for heads to come up 5 times in a row is 0.5 multiplied by 5, or 3.125%, we understand that the true value of the probability of something unlikely happening in front of you is on the whole less than 10%.

[REF] Multicollinearity
Multicollinearity is when there are more than two strongly correlated explanatory variables, which causes the calculated results for the partial regression coefficient to become unstable. There may be sign reversals, when the partial regression coefficient for a variable that would have had a positive correlation on its own instead becomes negative. We know that interpreting these results is difficult. One way of dealing with it is to exclude one of the explanatory variables with high correlation. However, if the purpose is only to forecast and it is not important to interpret the variable, there is no need to worry about multicollinearity.

Dummy variables

Some of the most important business decisions rely on qualitative variables instead of quantitative ones. Many of these variables cannot be expressed in numerical or monetary terms. So far, the analysis has looked at numerical data such as rent, living area, or age of a building. The result is a coefficient of determination adjusted for degrees of freedom of 0.91 and a fairly persuasive regression equation, but suppose it occurs to you that the direction the apartment faces (south-facing or not) may have an effect on the rent. How can we work this type of qualitative variable and categorical data into the regression equation?

We can actually use dummy variables to incorporate categorical data in multiple regression analysis. Dummy variables usually take the value 0 or 1. For example, if we introduce the new variable of south-facing, this is the format for converting it to quantitative data:

- Facing south (1)
- Not facing south (0)

What should we do if we want to make a rigorous distinction between the three categories of South-facing, North-facing, and Other? In this case, we prepare two dummy variables. The south-facing and north-facing variables are:

- South-facing (1, 0)
- North-facing (0, 1)
- Other (0, 0)

You may feel the urge to introduce the “other” variable, but this is not necessary as a zero for both variables corresponds to the “other” category. We always introduce one dummy variable fewer than the number of categories.

So, let’s look at the results of introducing the south-facing variable.

SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.9598
R Square	0.9211
Adjusted R Square	0.9126
Standard Error	10,099.9309
Observations	42

ANOVA

	df	SS	MS	F	Significance F
Regression	4.0	44,085,848,890.1	11,021,462,222.5	108.0	0.0
Residual	37.0	3,774,318,350.4	102,008,604.1		
Total	41.0	47,860,167,240.5			

	Coefficients	Standard Error	t Stat	P-value	Lower 95%	Upper 95%	Lower 95.0%	Upper 95.0%
Intercept	112,251.2	10,120.7	11.1	0.0000	91,744.8	132,757.6	91,744.8	132,757.6
Walking (min)	-2,991.1	533.5	-5.6	0.0000	-4,072.0	-1,910.3	-4,072.0	-1,910.3
Age (years)	-1,636.0	153.1	-10.7	0.0000	-1,946.2	-1,325.8	-1,946.2	-1,325.8
Exclusive living area (m2)	1,843.8	237.9	7.8	0.0000	1,361.8	2,325.8	1,361.8	2,325.8
South-facing	-571.2	3,918.1	-0.1	0.8849	-8,510.1	7,367.6	-8,510.1	7,367.6

With the addition of an explanatory variable, the coefficient of determination has increased insignificantly to 0.9211 from 0.9210 when compared to the previous model, but the coefficient of determination adjusted for degrees of freedom has decreased from 0.914 to 0.912. This means that the model where the south-facing variable was not used was a better model. In addition, the P value (risk factor) for the south-facing variable is 88%, which far exceeds the 10% criterion, so it is difficult to deny that this variable is actually zero (no significance). The idea was that the direction of the apartment would have an effect, but we understand from the results of the calculations for this example that adding the south-facing variable has almost no significance, rather the model without this variable is a better formula.

How to choose explanatory variables for multiple regression analysis

Considering the degree of difficulty of interpretation and explanation, the simplest possible model is more convenient. Broadly speaking, there are two methods for selecting explanatory variables.

1) The hypothesis-based approach
2) The search-based approach (automatic selection)

Method 1) involves building a hypothesis around causality and selecting variables based on repeated testing. The case of the rent for the studio apartment falls into this category. The explanatory power of the resulting regression equation may not be at maximum, but it is easy to understand and easy to explain the reasons for using that variable.

Method 2) is generally referred to as the stepwise procedure. You enter all the candidates that have potential as explanatory variables and then the software selects the optimum model based on defined standards. The explanatory power of the resulting regression equation is high, but it is difficult to explain why which variable were used.

Both methods have their advantages and disadvantages, but the automatic selection of variables is frequently used. Statistical software such as SAS or SPSS, or commercial add-in software for Excel, such as Excel Analysis Add-in usually have a function for automatic selection of explanatory variables, and will select the variables automatically.

Unfortunately, Excel's basic version does not have an automatic function for selecting this variable, but it is possible to manually select pseudo variables. This is backward stepwise regression that removes variables from a model where all variables have been entered.

- Automatic variable selection using stepwise regression (assuming Excel)
  1) Select groups of candidates for the objective variable and the explanatory variables based on the hypothesis.
  2) Narrow down the explanatory variables using backward elimination.
  - Use the analytical tool for correlation to create a correlation matrix. To prevent multicollinearity, one of the variables with a correlation efficient above 0.9 is removed from the explanatory variables at this stage. (Normally, the causal series remains.)
  - Use all the explanatory variables to perform the regression analysis with the analysis tool.
  - Remove the explanatory variables with the highest P value from the results and build the regression equation. Repeat until you have only one explanatory variable.
  - Choose the model with the highest adjusted R-square.
  - Confirm: Is the P value for each explanatory variable in the model smaller than 10%?

Data on studio apartments in the vicinity of station A

Walking (minutes)	Age (years)	Living space (m^2)	South-facing	Rent + common services charge
13	24	22.15	1	69,000
6	33	17.83	0	70,000
5	31	23.63	1	80,000
9	28	27.76	1	84,000
7	36	23.32	0	87,000
12	2	31.03	0	116,000
7	12	26.22	0	119,000
4	6	22.16	0	129,000
1	7	36.54	0	166,000
4	6	40.04	0	170,000
7	33	17.83	0	70,000
8	35	23.1	0	76,000
2	41	16.11	0	69,000
13	18	17	0	72,000
5	31	23.63	1	80,000
8	35	25.8	0	82,000
6	28	21.8	1	75,000
4	34	22.33	0	79,000
2	41	18.27	1	83,000
5	29	27.06	0	85,000
8	29	26.7	0	85,000
6	28	29.06	0	98,010
11	26	20.85	0	83,000
11	35	21.31	0	84,000
5	30	29.52	0	96,000
7	30	34.41	0	94,000
5	30	35.79	0	108,000
4	34	30.82	1	105,000
4	33	30.82	0	110,000
9	23	33.78	1	115,000
9	23	34.02	0	118,000
9	9	32.76	0	125,000
4	6	22.16	0	129,000
10	27	44.89	0	130,000
6	7	29.39	0	142,000
6	7	47.97	0	149,000
6	7	32.71	0	154,000
6	9	36.82	0	154,000
1	7	36.54	0	158,000
4	6	40.04	0	170,000
5	6	42.82	0	178,000
4	6	30.2	1	169,000

The inductive approach objective is to infer whether or not a relationship exists among the variables and then to extract a reproductive mechanism which has significant influence on the outcome. Be careful not to confuse strong correlation as implying the existence of causation.

(B) The Deductive Approach

If you have been to an interview with a consulting firm or an investment banks, you may have been asked some "brain-teasers," such as guessing the number of piano tuners in Chicago, the number of telephone poles in Japan, or the number of new cars sold annually in Japan.

One commonly known deductive approach is the Fermi estimation. Different from the inductive approach, which tries to estimate appropriate relationships from existing data, the deductive approach is an estimate based on approximate calculations from limited data. This is when creating a logical and realistic model can become useful, for situations which are impossible to measure.

Modeling

Formulate relationships through regression analysis is an approach based on actual data that inductively describes the relationships behind the data. Modeling, on the other hand, is an approach that uses deduction to formulate the relationship between the inputs and outputs. With modeling, the interview questions mentioned above can be easily answered with some guesses and simple calculation.

Broadly speaking, in terms of outputs, there are models that seek stock (the number of vehicles owned etc.) and models that seek flow (sales etc.).

Stock series: Number of vehicles owned by Japanese families = Number of households × Vehicle ownership ratio × Number of vehicles per household
Flow series: Café sales = Number of seats × Occupancy rate × Turnover × Average spend per customer

The modeling approach that tries to find simple ways to understand seemingly complicated phenomena is extremely versatile. For example, it can be applied to the following:

1) To understand business mechanisms and earnings structures from multiple perspectives
→ Confirm and discover the structure or features of the business in question
2) Forecasting and sensitivity analysis
→ Distribution of business resources, risk management, rebuilding business models

- A word of caution about modeling

It is almost impossible to survey all numerical values when you want to find out the range of figures necessary for a business. Surveys also take up a great deal of money and time. It is only wise to think about how you can break down the figures you want to study into their elements, quantify them, and make estimates to the extent possible.

If possible, try to create several other models for cross-checking and estimates.

For example, if you want to estimate sales, you can create equations using the range of perspectives outlined below.

Sales = Sales to existing customers + Net sales to new customers
Sales = Number of customers × Average purchase amount per customer
Sales = Industry sales × Your company's share of the market
Sales = Fixed costs + Variable costs + Operating income
Sales = Average daily sales × Number of days
Sales = Actual sale for previous fiscal year × Growth rate
And so on…

In addition, if you combine several of the above approaches, you will be able to create more detailed and comprehensive equations. But what are the most important factors should you consider when selecting a model equation for business use from among the countless conceivable equations? They are not all-purpose equations that will fit any business, and you need consider each one individually while keeping your circumstances in mind. These are some points that you should keep in mind.

(1) When modeling a business you know

Observations are key, so day-to-day observations of what elements will work well for your business are important.

Example: Beer/Temperature, Houses built/Interest rates, Urban drugstores/Number of passersby

Discovering the most important figures for your business through careful, in-depth study is valuable, even if the data is only shared across the company.

(2) When modeling a new business

Rather than fussing over the subtlety and uniqueness of the model, your first priority should be to perform a rough analysis using addition, multiplication, and division.

Example: How many customers can you expect if you start up an educational business?
Registered applicants = Number of inquiries × Attendance ratio at orientation × applicant ratio + …

To make a logical business decision, it is necessary to determine a model, evaluate the alternatives using that model, and select the best solution. A model should be viewed as roadmap and a simplification of the real problem which incorporates the essence of your business. More importantly, it allows you to identify the existing strengths or weaknesses of your business; thus, you may be able to manage the resources and the risks more effectively.

Unit 4: Decision-Making under Conditions of Uncertainty

Unit 4-1: Uncertainty in Decision-Making for Businesses

Business is nothing less than the continuous pursuit and implementation of methods to realize the goals that you have set for your business. In Unit 3, we saw that it is essential to have an understanding of causality for the purpose of achieving the goals and selecting the appropriate methods.

However, in business, you may encounter situations where causality is not well understood to begin with, or, even if causality is understood, you are often confronted with situations where you cannot escape uncertain outcomes caused by fluctuating input factors, such as exchange rate fluctuations that have an impact on earnings in the export industry.

We will study suggestions for how to approach decision-making under such conditions of uncertainty (when it is not possible to predict what will happen in the future).

Uncertainty and risk

In any business field, especially in finance, uncertainty is normally expressed as risk. In this context, risk refers to uncertainties that are unpredictable.

To start with, it should be noted that the word risk is used differently in business than in ordinary settings. In everyday language, the word risk is frequently used where you encounter some danger, or there is possibility of injury. In business, the word risk expresses dispersion and uncertainty in the outcome regardless of whether the outcome is good or bad.

For example, suppose we all play cards together. Let's assume that if you lose, you have to pay a penalty. You may win and avoid the penalty, or you may be unlucky and lose, and then you have to pay the penalty. During the game, whether or not you can avoid the penalty is uncertain and the risk is high. Incidentally, will there be any risk after you lost the game? In everyday usage, it may seem as if the risk has peaked. However, in business usage, if the penalty becomes a certainty, the risk is zero.

How to confront risk

In terms of dealing with risk, we will learn the two items below in the following section:

1) Sensitivity analysis to ascertain risk through visualization

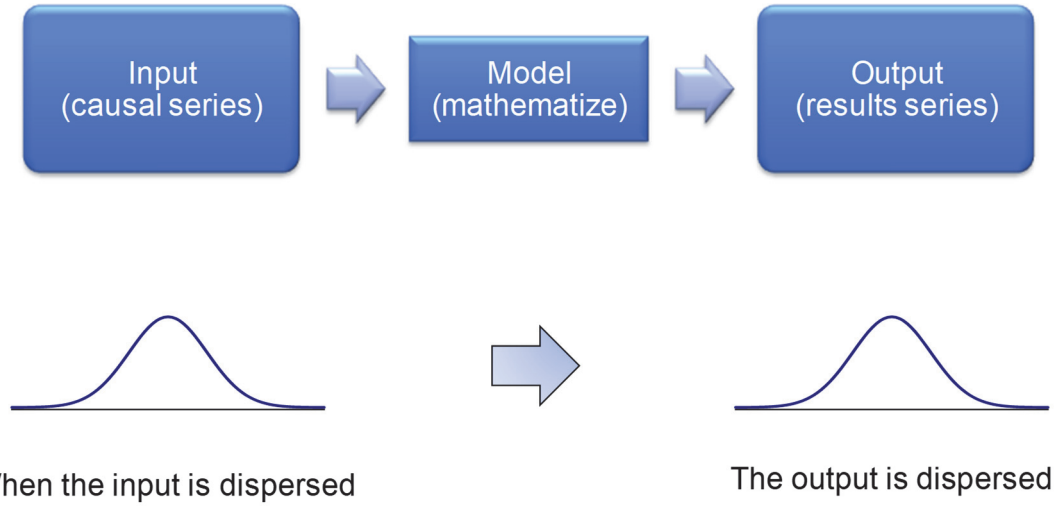
2) The decision tree, which is a way of picking options when risk (uncertainty) is present

Unit 4-2: Sensitivity Analysis (Tornado Charts)

Sensitivity analysis using tornado charts is a method for investigating feasible actions based on the visualization of risk sources and the size of the risk. With modeling, we learned to understand the relationship between outputs and inputs and to use deduction to analyze and formulize a business.

In sensitivity analysis, we make use of formulas created for modeling. This is a technique for quantitatively understanding the degree of uncertainty in the input, and, as a result, to what degree the risk and uncertainty occurring in the output is caused by the input. Then, the technique can be used to search out improvement trends in order to reduce risk.

In order to visualize risk, we will discuss the so-called Tornado Chart at the end of this section.



The following are the steps in sensitivity analysis.

1) Modeling

Structure your company's business, then model (mathematize the relationship between input and output) the variables and parameters and how they are connected.

2) Analysis of the level of sensitivity

(1) For each variable, quantify the degree of impact on the output when the variable is activated
(2) Visualize the result in a tornado chart
(3) Study the quantified values and identify particularly important factors by ranking the variables

3) Ascertain and deal with risk

By taking measures to deal with the points that you identified as particularly important, you will be able to create a risk-resistant business plan.

Let us look at some examples.

Suppose you are a taxi driver for a taxi company. You would like to start your own taxi service but there are some legal conditions since you are less than 35 years old, and

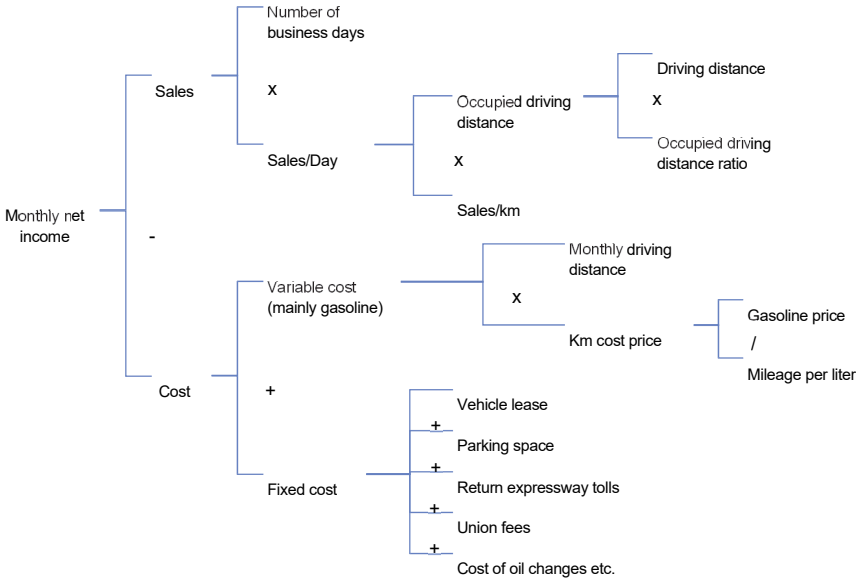
- You have worked at least 10 years continuously as a driver for a taxi company,
- You have a clean driving record for 10 years before the date of application.

Now, you met these requirements, at the point of switching to a privately owned taxi business.

You are worried about income and expenditure when you start to operate the privately owned taxi business. In order to pay the mortgage and the school tuition for your children, you need to make at least 350,000 yen per month before taxes. Since gasoline prices are also fluctuating, you are not certain you will take home at least 300,000 yen after expenses and taxes are taken out.

Step 1: Modeling

First, you try to model income and expenditure based on information collected from a colleague who already operates a privately owned taxi.



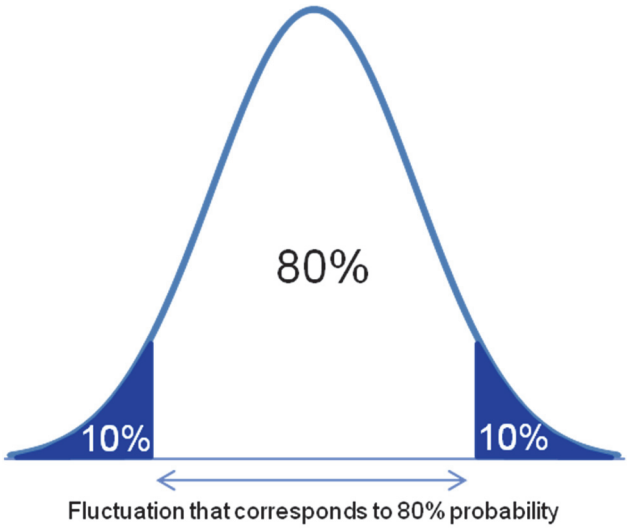
＊Driving distance: Distance actually driven; Occupied driving distance: Distance driven with passengers in the taxi; Occupied driving distance ratio: Occupied driving distance/Driving distance

Step 2: Sensitivity analysis

Based on the performance of the more experienced taxi driver, the following factors vary little:

- Number of working days: 20 days
- Sales/km: ¥400
- Mileage: 10 km per liter
- Fixed cost: ¥185,000

On the other hand, driving distance, occupied driving distance and gasoline prices seem to vary a great deal. You will have to consider to what degree they are likely to vary.
Sensitivity analysis using tornado charts excludes the top and bottom 10% probability in the fluctuation band, and analyzes the range of fluctuations where probability is assumed to be 80%. There is no statistical evidence for excluding the 10%; it is simply that 10% is the criterion used for unlikely events.



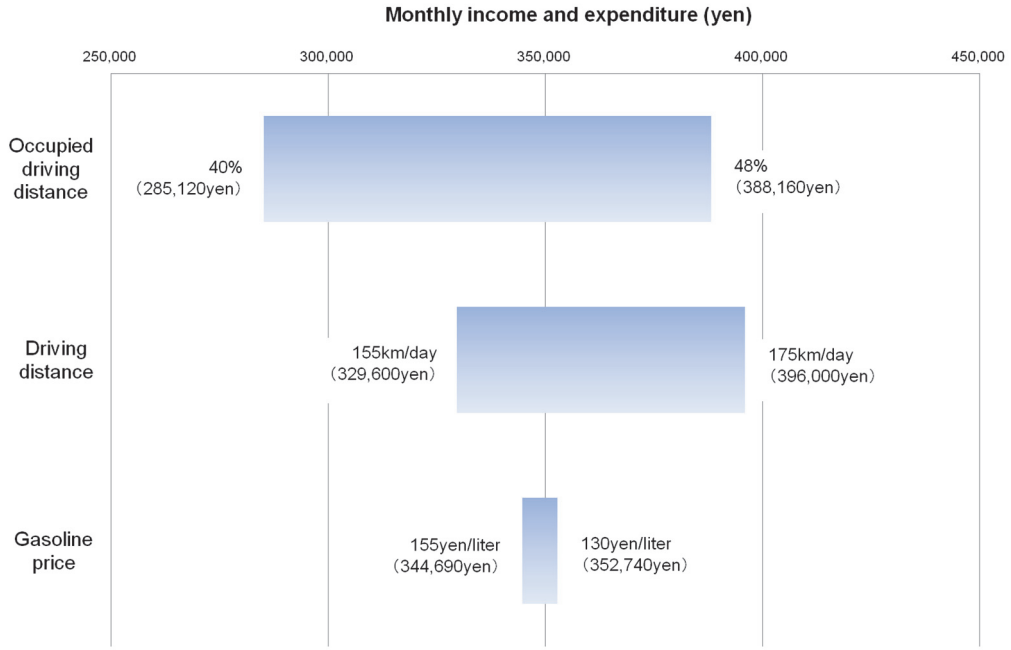
These fluctuation estimates are based on the performance and past data provided by the more experienced driver:

	High scenario	Base scenario	Low scenario
Driving distance (km/day)	155	161	175
Occupied driving distance (%)	48%	45%	40%
Gasoline price (yen/liter)	130	140	155

Calculations based on this data show the following monthly income and expenditure for the above scenario:

	Driving distance (km/day)	Net income
High scenario	175	¥396,000
Base scenario	161	¥349,520
Low Scenario	155	¥329,600
	Gasoline price (yen/liter)	**Net income**
High scenario	130	¥352,740
Base scenario	140	¥349,520
Low Scenario	155	¥344,690
	Occupied driving distance (%)	**Net income**
High scenario	48%	¥388,160
Base scenario	45%	¥349,520
Low Scenario	40%	¥285,120

All the numbers may look confusing and provide little value if we are not looking carefully. Let' s try entering the data in a tornado chart. The tornado chart earned its name from its shape, which resembles a tornado since the data is normally entered in descending order from the variables with the highest range of uncertainty.



Step 3: Ascertain and deal with risk

From the tornado chart, we understand that, in relative terms, a major risk factor is not gasoline prices, but rather occupied driving distance.

In particular, if the occupied driving distance drops down to 40%, income and expenditure look set to fall below the 300,000 yen limit, which is a concern. We have to think about what can be done to make sure the occupied driving distance does not decrease in the first place.

You need to figure out ways to catch up with the demand, for example, by carefully reading the daily event listings for the whole city, and not wait around.

Unit 4-3: The Decision Tree and Uncertainty

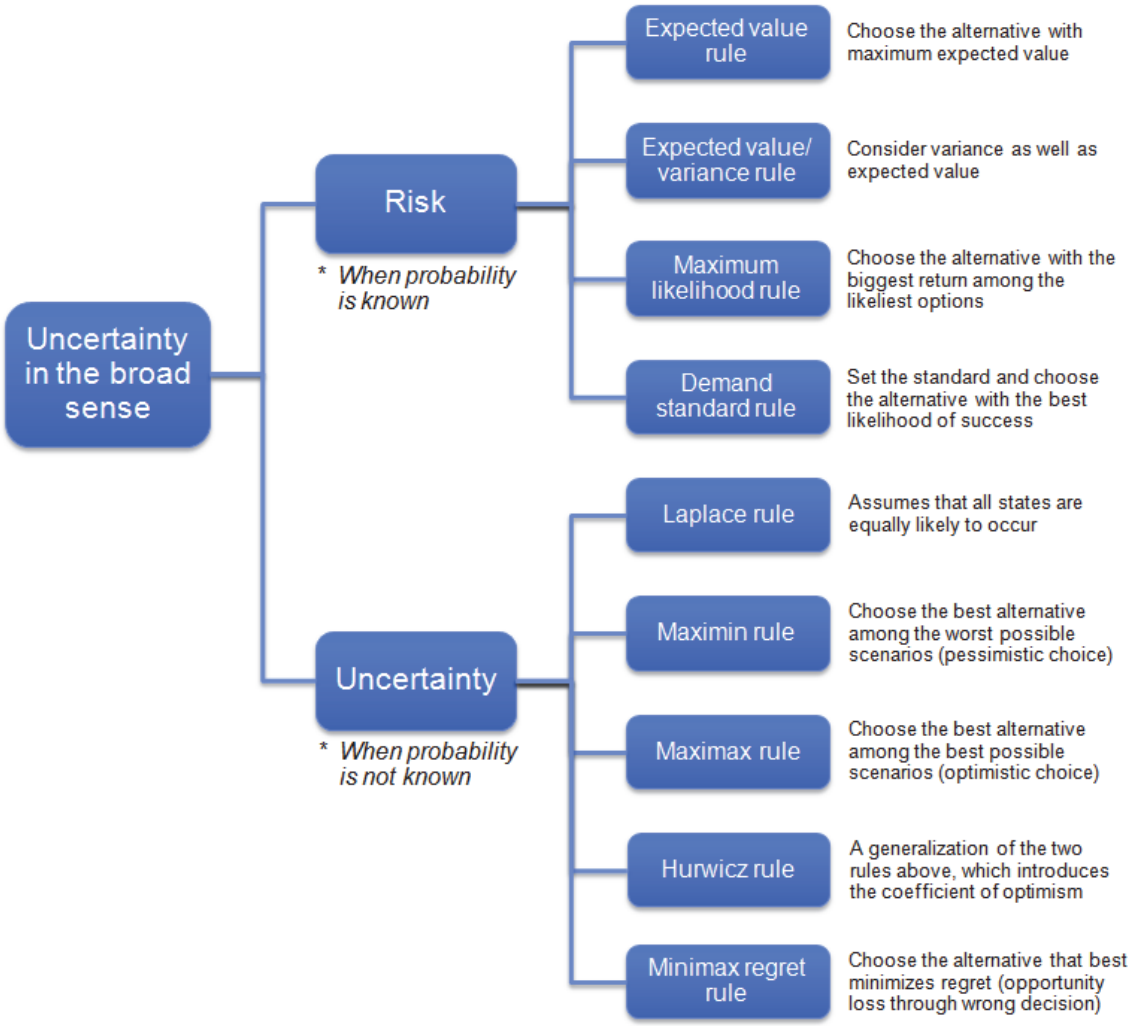
Selecting alternatives under uncertain conditions

Managers are not only required to process or interpret events that actually occurred in the past, but also to make decisions by comparing and analyzing the uncertainty in future events.

In cases where you are given options, there are several alternatives for evaluation criterions under uncertain conditions as outlined in the diagram below. The expected value rule at the top corresponds to the decision tree discussed in this chapter, while the second rule, the expected value/variance rule, applies to the portfolio concept in finance.

* We already explained that uncertainty almost equals risk, but this classification provides a more detailed breakdown where risk in the narrow sense is when you know the probability of something occurring, and uncertainty is when you do not know the probability.

Here we will discuss the typical decision tree, which corresponds to the expected value rule at the top.



The decision tree approach

The dispersion of uncertain future events can be understood as probability.

- Probability is a value between 0 and 1 that can be assigned to a particular event
- The sum of the probability of all events is 1
- If two events are mutually exclusive, the probability that either of the events happens is the sum of the probability that each happens

The decision tree visualizes and applies this approach to probability.

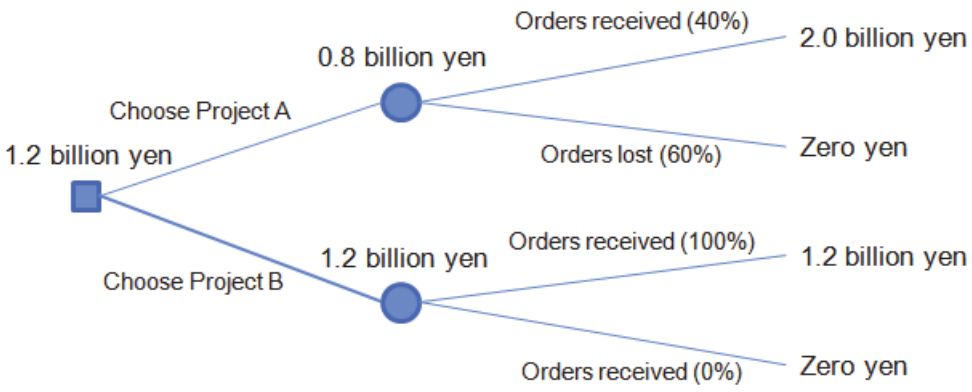
The decision tree is a tree-shaped diagram for expressing combinations of alternative decisions and the uncertain events that have impact on the alternatives. The following are the benefits of drawing a decision tree:

1) You gain a chronological understanding of which uncertain events should be considered for each alternative, and the details of decision-making.
2) You can compare uncertain alternatives by using expected values

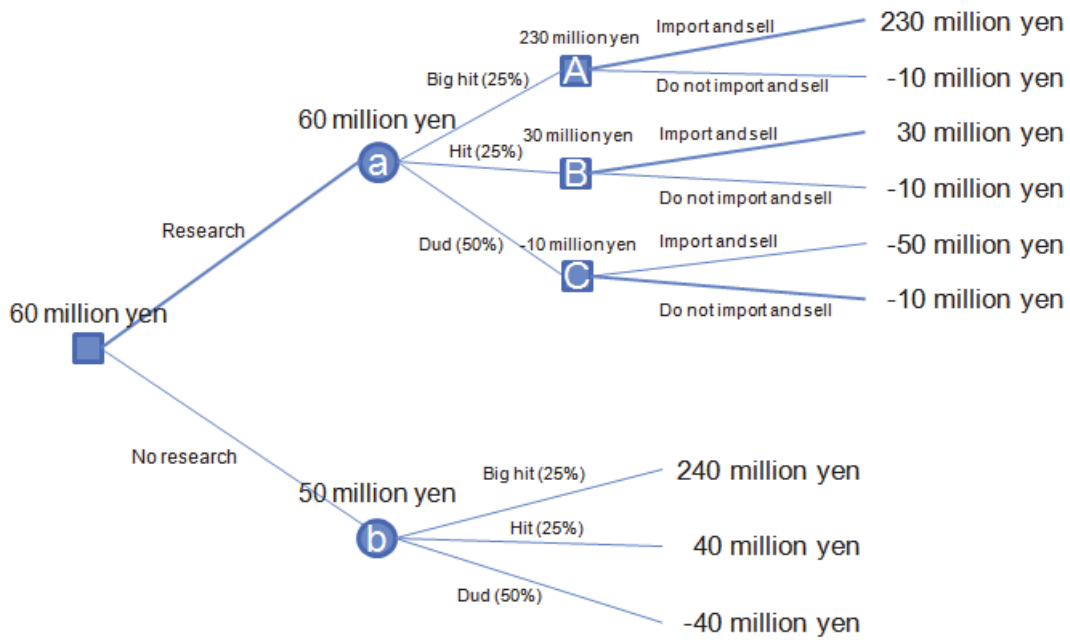
Generally, you draw a decision tree starting from the base and working out toward the branches, noting the alternatives in chronological order. Square nodes represent decisions and any branches from the nodes indicate alternatives for decision-making. Round nodes indicate new information and any branches indicate scenarios that may develop.

For example, in the diagram below, the square node on the far left represents the decision to go with project A or project B. The adjacent round nodes marked 800 million and 1.2 billion indicate the expected value of orders received after selecting project A or B. The branches from 800 million and 1.2 billion show possible scenarios and their probability.

If, for argument's sake, we rely solely on these figures to make a judgment, we understand that it would be more rational to choose project B (1.2 billion yen), which has a higher expected value than project A. Consequently, we start by writing the expected value for project B in the square node.



Next, we will consider a complex decision-making process in order to develop a deeper understanding of the decision tree. Assume that we are considering whether or not to import and sell products from overseas. Three scenarios are conceivable: the product becomes a big hit (profit is 240 million yen) / it becomes an ordinary hit (profit of 40 million yen) / it ends up as a dud (loss of 40 million yen) From past experience, we know that the probabilities are 25%, 25% and 50% respectively. We also have a choice of whether or not to first spend 10 million yen on research. Depending on the result of the research, we have the option to choose not to import and sell the products. If we do the research, we will definitely know if it will be a big hit, an ordinary hit or a dud. However, we will not know the outcome of the research in advance. Therefore, let's assume that the research results will also be 25%, 25%, 50% probability of being a big hit, an ordinary hit or a dud, respectively.



As you can see from the diagram, in cases where decision-making involves multiple stages, we usually determine the optimum choice at each node, working from the end of the branches to the base of the tree (on the left). Then we can determine the difference between the expected values at the node that is closest to the base of the tree.

For example, at the decision-making node labeled A (top right in the diagram), the alternatives are "Import and sell" and "Do not import and sell," and if we make the judgment based on the figures, we will start to import and sell the product. Similarly, we should choose the alternative to sell at node B, and the alternative not to sell at node C.

As a result, if we do the research, the expected value is:

> 230 million × 25% + 30 million × 25% -10 million × 50% = 60 million yen (value at node a in the diagram).

Similarly, if we do no research, the expected value of importing and selling is:

> 240 million × 25% + 40 million yen × 25% - 40 million yen × 50% = 50 million yen (value at node b in the diagram)

Based on the above, the round node A yields 60 million yen and the round node b yields 50 million yen. So based on these figures, we can conclude that conducting research is rational.

Also, from the difference in expected value (10 million yen) between when you do research and when you do not research, we can calculate the value of research: current cost of 10 million yen + 10 million yen difference of expected value = 20 million yen. Hence, the value of the research information is maximum 20 million yen. In other words, the answer to the question as to how much you should spend on research is that you can expect additional profits if you spend under 20 million yen.

Where does the value of the research as information originate? If we look at the tree, we see that whether the research was done or not, the action changes if the product is a dud. If you do the research and realize that the product is a dud, you will decide against importing and selling the product, but if you do no research, the action changes because you will import and sell no matter what. The action changes depending on the information you have obtained and as a result, the expected value changes. This difference is the value of the information.

So far, we have been considering the expected value based on the best choice under the assumption that we know the probability. However in real business, it is more likely that we do not know the probability. What if we do not know the probability? The following shows the common approaches, as described in P.71, when the probability is not known.

1. Laplace rule
   - Assume that all states are equally likely to occur
   - Expected value of doing the research is 83.33 million yen, while the expected value of not doing the research is 80 million yen. In this case, we would choose to do the research.
2. Maximin rule
   - Choose the best alternative among the worst possible scenarios
   - The worst scenario if we do the research results in 10 million yen loss which corresponds to square node C, while the worst scenario if we do not do the research is 40 million yen loss. In this case, we choose to do the research.
3. Maximax rule
   - Choose the best alternative among the best possible scenarios
   - The best scenario if we do the research results in 230 million yen profit, while the best scenario if we do not do the research yields 240 million yen profit. In this case, we choose not to do the research.
4. Minimax regret rule
   - Choose the alternative that minimizes regret, where regret is defined as the opportunity loss through having made the wrong decision.
   - The expected values of the 3 scenarios if we do the research are 230 million for big hit, 30 million yen for ordinary hit and -10 million yen for dud. If we do not do the research, the expected values are 240 million yen for big hit, 40 million yen for ordinary hit and -40 million yen for dud. Hence, the largest regret if we choose to do the research among the 3 scenarios (240-230=10 million yen, 40-30=10 million yen, -40-(-10)=-30 million yen) is 10 million yen. On the other hand, the largest regret if we do not do the research among the 3 scenarios (230-240=-10 million yen, 30-40=-10 million yen, 40-10=30 million yen) is 30 million yen. In this case, we choose to do the research, which has the smaller of the regrets.

We have seen the various approaches for making decisions. Even if the same information is given, the decision can be different depending on the criteria.

Which decision criterion should you choose after all? Unfortunately, there is no principle for choosing decision criteria. Decision makers have to decide for themselves. However, decision-making involves accountability, so it is necessary to be able to explicitly explain the reason of your decision.

Column: Big Data, Machine Learning and Artificial Intelligence

You must have heard of the term, Big Data. Incidentally, by examining the search frequency of the word on Google Trend, we see that its usage started to grow rapidly from around 2011 (actually this search data itself is big data).

Let us first look at the differences between the features of "big" data and the traditional "small" data.

Just having big data and not doing anything with it, it would be a liability instead of an asset because of its high costs. Analysis is necessary to convert big data into value. We will briefly explain machine learning which is an analytical method necessary to extract value from big data in business, and also artificial intelligence (deep learning) which is a type of machine learning.

Difference between big data and small data

There are several definitions as to what big data primarily refers to. The most famous one is probably the 3Vs defined by Gartner Inc., a US research company.

Volume:	Huge amount of data
Velocity:	High speed of data flow. Data is frequently updated like the GPS location data.
Variety:	Various types of data (not only conventional quantitative data but also data such as text on social media, images, movies, sounds etc.)

The amount of data that we deal with is exponentially increasing. In 2010, Eric Schmidt, then Google CEO, symbolically said that the amount of information that humans produced in the past two days is equivalent to the amount of information that has been produced from the dawn of human civilization to year 2003. Take social media for example, Twitter generates 7.7 thousand tweets per second, which sums up to 660 thousand tweets in one day. 800 photos every second or 69 million photos per day are uploaded to Instagram. It is also known that Google's search volume is 61,000 per second, or 5.3 billion cases per day[5].

In addition, the amount of data acquired from sensors has increased dramatically as they become cheaper and more accurate. For example, GPS location data is continuously being collected from smartphones and is being updated in real time. With the progress of IoT (Internet of Things), it is said that 100 trillion sensors will be used by 2023, resulting in huge growth in the amount of data collected in real time. IoT and Big Data go hand in hand.

Other than the growth in data quantity, the change in data quality has also contributed to the dramatic improvements in the quality of data analysis.

The purpose of collecting data in business is simply an attempt to quantify the real world and use it to achieve business objectives. In other words, we are trying to represent the real world with data. For instance, it is extremely important to understand our customers in order to have them purchase our goods and services. Conventionally, we could only gather information from

[5] http://www.internetlivestats.com (Accessed 26 July, 2017)

limited sources such as customer questionnaires and actual sales, in addition to knowing the customer attributes.

With the advent of big data (such as social media posts and behavioral data through GPS sensors), we can now quantify customer behavior and thoughts that were impossible to collect in the past. Thanks to big data, it is now possible and becoming very important to analyze and understand the actions of individual customers and utilize the information for business.

Leveraging Big Data with Machine Learning

Let us examine the differences between big data and small data from a different angle from the 3Vs: how to effectively use the data.

Before the era of big data, data was limited, and we tended to collect all the data we could acquire and use all of them in analysis. It would take huge amount of effort (time and money) to collect large amounts of data. In Japan for example, the national census survey carried out every five years collects data from all citizens in Japan (in that regard it could be considered as a predecessor of "big data"). However, doing so requires a huge number of investigators, and enormous amounts of time, labor and cost. In 2015, the census comprised of about 700,000 researchers and a budget of 67.2 billion yen.

Since the cost of data acquisition before the big data era was extremely high, instead of collecting all data, the common approach was to gather a portion of the data at low cost (sampling), and to try and estimate the whole data population from the data samples. Hence, big data like the census was rather exceptional. An example of data sampling is the survey of the Japanese Cabinet approval rate by using opinion polls, in order to predict election results. Inferential statistics which estimate the whole from small sample sizes was the star of statistical studies. It was one of the most important fields of studies in modern statistics.

By examining data in its whole, we can get high level of accuracy, but in the days when data acquisition cost and analysis cost were prohibitively high, data was a scarce resource.

However, as previously depicted in 3Vs, the era in which it is common for us to have large amounts of data has come. In the case of big data, we consider the whole data rather relying on a small set of samples. This characteristic has dramatically shifted the position of data from being a scarce resource to becoming an abundant resource. Along with that, data analysis has changed; not only the scale of analysis, but also the analytical method itself.

In the world of big data, analytical method to extract patterns from data, called machine learning, has emerged, taking over the incumbent method of inferential statistics.

The value of big data lies in the extraction of patterns hidden in the data. Without machine learning, data is just a mountain of useless information. Deep learning, which is often discussed in artificial intelligence, is also a type of machine learning. We can also associate machine learning with the field of data mining almost synonymously. In fact strictly speaking, machine learning is not necessarily artificial intelligence, but in recent media, these terms are used synonymously.

Actually, regression analysis that we learnt in this course is also one of the machine learning methods.

Machine Learning cannot explain why

Machine learning is extremely powerful when it comes to making use of big data, but there is actually a big weakness when compared with conventional data analysis. It predicts and classifies big data with high precision, but it does not explain why.

”Machine learning is not good at explanation”.

For example, it is possible to use machine learning to estimate emotions, e.g. smile or frown, from facial images. But we do not know why the machine learning categorizes this way and what criteria it is using. Even if you look into the mechanism of machine learning, it is often too complicated for human beings to understand. It might not be correct to say that conventional analytical methods can explain why, but since data volume is small and analysis is simple, human beings can easily interpret the results causally (“this is how it seems to be”). Machine learning of big data makes causal interpretation extremely difficult. Some people regard this as the characteristic of big data, that ”correlation (pattern)” is more important than ”causality” in big data.

Human beings tend to always seek for causal explanations. However, in the case of recommendation engine in an online shore where machine learning is used, the reason why a particular product is recommended is not clear, and yet no consumers are asking for explanations. If you like the recommendation, you buy it, if not, you do not buy it. Machine learning can predict the items you may buy from your purchase history, and as long as the product is actually purchased, it does not matter in business even if the recommendation cannot be explained.

Big data and the role of human beings

We have seen the features of big data and how value can be created using machine learning. What should we then keep in mind when using big data? Machine learning is indispensable for utilizing big data, but as we have discussed before, it is weak at explaining why.

When making business decisions, it seems that conventional analytical methods will still be the mainstream for higher order decision-making, such as strategic decisions, which requires causal explanations. Meanwhile, big data and machine learning will become the mainstream in areas where explanations are unnecessary, and in areas that can be routinely automated like the recommendation engines. However, it also seems that the need for explanation is changing with times. Humans tend to seek for explanations, but at the same time, they also tend to adapt well to changing environment. Even though we might first seek for explanations, but as we continue to achieve reliable prediction results through Big Data + Machine Learning, we might eventually stop asking for further explanations.

Machine learning does not collect data on its own. Unless humans define the purpose/problem and collect the actual data, analysis cannot happen. In other words, upstream processes of big data analysis must continue to be handled by human beings.

The ability to grasp the purpose and issue, and to deduce what kind of data is necessary is more important than ever in the era of big data.

<Problems>

Problem1: Difficulties in Quantitative Analysis

Describe situations where things did not go well in the past when you used figures for analysis and reported the results to someone. What were the reasons things did not go well?

Please consider the question carefully and give specific examples.

* You are not required to submit this problem.

Problem 2-1: Inadequate Questioning

The following questions used in a questionnaire are inadequate. Explain why they are inadequate.

1) How many times do you go shopping or eat out in a month?

2) Choose the statement which is the closest to your view regarding a government official leaving for a high position in a private company.

- Should be stopped immediately.
- Each case should be considered individually.
- Maintain as it is.

3) Some people say, "Studying for entrance exams is bad because it takes away freedom from children." Do you agree or disagree with this view?

4) Are you satisfied with the design of your mobile phone?

5) How many close friends do you have whom you can talk to when you are in trouble?

6) The question: "Do you think Japan should contribute to the international community in accordance with its current national strength?" This is followed by the question: "Do you think Japan's Self-Defense Forces should be dispatched for international use?"

Problem 3-1: Apple Tree Fitness

Create a chart based on the case below, and draw an inference from the data.

Background

Apple Tree Fitness is an independent fitness club in Tokyo located near a private railway line. The club mainly targets women, but recently the number of male members has increased. Business results have been improving every year, although slowly.

As the owner, you are concerned that the customers' age may be shifting upwards, based on your observation during a site visit. So far, there does not seem to be much impact on the profit ratio, etc., but if customers are aging, it could impact the business in the long run.

Apple Tree Fitness normally asks customers to provide their address and contact details when they became members, but has a policy of not asking for information on their age or marital status. Therefore, to find out their age, the company sent a questionnaire to customers who volunteered, and offered them a small reward for participating.

The following are the results of the surveys done both three years ago and this year.

Ages of customers: Three years ago

18	19	21	21	22	23	23	24	26	26	26	26	27	27	27	28	28	29
30	30	30	31	31	32	32	33	34	35	35	36	38	38	39	40	41	42
42	42	43	43	44	44	44	45	46	48	48	48	48	52				

Ages of customers: This year

21	22	23	24	25	26	26	27	27	27	27	28	28	28	29	30	31	31
31	32	32	33	34	35	36	36	37	39	39	40	40	40	41	43	43	43
44	44	45	45	45	45	46	46	47	48	48	49	51	55				

Problem 3-2: Quantitative Survey of Convenience Stores

The table below shows fluctuations in the number of convenience stores nationwide from 2008 to 2011 based on a quantitative survey of convenience stores.

Try to draw a graph. What interpretations can you draw from the graph?

Right-click on the data in the graph, select Add Trendline > Linear and use the linear expression to try to get the approximation formula that indicates trends. Does it seem possible to tell what the trend is like from the equation? Can you predict from the equation how many stores there will be in December 2012?

Year	Period	Number of stores	Year	Period	Number of stores
2008	January	40,889	2010	January	42,704
	February	41,105		February	42,919
	March	41,204		March	42,815
	April	41,356		April	42,865
	May	41,398		May	42,879
	June	41,367		June	42,889
	July	41,443		July	42,995
	August	41,643		August	43,270
	September	41,566		September	43,281
	October	41,559		October	43,268
	November	41,666		November	43,291
	December	41,714		December	43,372
2009	January	41,800	2011	January	43,393
	February	42,047		February	43,636
	March	42,004		March	43,492
	April	42,070		April	43,492
	May	42,153		May	43,560
	June	42,204		June	43,541
	July	42,345		July	43,690
	August	42,557		August	43,985
	September	42,487		September	43,969
	October	42,553		October	44,062
	November	42,673		November	44,250
	December	42,629		December	44,403

Source: Japan Franchise Association

Problem 3-3: Right Decision

Setting a Target

Read the case below, and determine the annual target figures for each item for the Osaka area that should be achieved by the end of three years from now. These target figures will be used to appraise performance of the managers.

Background

Right Decision is a Tokyo-based consulting firm and provides services to middle-sized companies. All the firm's clients are currently based in the vicinity of Tokyo, but the firm is now considering expanding its business into the Osaka area.

At the moment, there are four teams working for the Tokyo area. These teams were formed based on the chemistry between the partner, who leads the team and is engaged in most important sales activities, and the team members. They are not based on industry or geography.

The Osaka area has a potential market of roughly a quarter to a third of the size of the Tokyo market. Three years from now, the firm wants to have an office of about 10 staff and achieve productivity and profitability comparable to the Tokyo area.

	Tokyo Team 1	Tokyo Team 2	Tokyo Team 3	Tokyo Team 4
Number of Staff	10	12	9	6
Number of Business Leads	175	222	119	130
Number of Proposals Submitted	113	135	77	65
Average Proposal Amount (million yen)	14	8	16.5	12.5
Number of Orders Received	25	45	18	20
Average Order Amount (million yen)	11.5	6.6	14.5	11.8
Total Amount of Orders (million yen)	287.5	297	261	236

- Items requiring target-setting (Use the same unit as in the table above.)
  Number of business leads
  Number of proposals submitted
  Average proposal amount
  Number of orders received
  Average order amount
  Total amount of orders

Problem 3-4: Management Selection Test

Examine the distribution of the data in the case below, and draw an inference. Based on your inference, consider what actions you would take if you were a corporate manager.

Background

Clearbiz has a policy of conducting an objective test which measures business framework, critical thinking and professional knowledge for selecting its managerial staff. The company only promotes candidates who score 60 points or higher.

The table below shows the test results and the performance rating (standardized scores representing the degree of achievements by those same individuals) three years after promotion of 42 randomly sampled managerial staff members.

Test Score At Promotion	Performance rating 3 years after promotion	Test Score At Promotion	Performance rating 3 years after promotion	Test Score At Promotion	Performance rating 3 years after promotion
60	90	70	59	81	86
60	45	72	78	82	90
61	60	73	78	84	58
61	66	73	63	85	70
62	55	74	91	86	72
63	52	75	73	88	75
63	88	75	70	88	81
64	60	75	80	90	77
66	66	75	94	92	63
66	62	77	81	92	86
67	90	78	74	92	84
68	72	78	92	92	97
69	64	80	99	95	99
70	70	81	84	95	90

Problem 3-5: ProSol

Analyze the correlations between the items in the case below, and draw an inference.

Background

Pro Sol provides consulting and development services in information systems. The company has annual sales of roughly 1 billion yen. As the company reorganized the data relating to its client companies as part of its efforts to increase sales, it compiled the following data.

Current Annual Sales (Unit: Transaction amount = thousand yen, Profit ratio = %, Transaction period = years)

Client	Transaction amount	Profit ratio	Transaction period	Client	Transaction amount	Profit ratio	Transaction period
1	2000	8	1	31	7300	12.5	4
2	2500	12	2	32	12000	15.8	7
3	3200	12.4	3	33	13000	16.2	6
4	4500	12.5	3	34	15600	14.5	4
5	5000	10.5	2	35	4000	8.7	2
6	2000	9.2	2	36	11050	9	1
7	1000	7.5	2	37	3500	7.7	1
8	500	7.6	3	38	4500	11.5	2
9	600	6.3	3	39	5500	11.9	3
10	1200	12.3	5	40	7500	12	3
11	1800	10.9	4	41	8500	10.1	2
12	2300	12.4	4	42	4000	8.8	2
13	5600	12.8	4	43	2000	7.2	2
14	9000	16.5	7	44	900	7.3	3
15	10000	16.5	6	45	1000	6.1	3
16	12000	14.8	4	46	2000	11.8	5
17	3000	8.9	2	47	2500	11.5	2
18	8500	9.2	1	48	2400	10	2
19	2600	7.8	1	49	3000	10.2	2
20	3250	11.8	2	50	3200	10.5	2
21	4160	12.2	3	51	1800	10.3	2
22	5850	12.3	3	52	1200	9	2
23	6500	10.3	2	53	1000	8	2
24	2600	9	2	54	500	7	2
25	1300	7.4	2	55	800	7.6	2
26	650	7.4	3	56	1500	8.5	2
27	780	6.2	3	57	1600	9.2	2
28	1560	12.1	5	58	1900	10.3	2
29	2300	10.7	4	59	2300	9.8	2
30	3000	12.2	4	60	2700	8.5	2

Problem 3-6: Company L / Company M / Company N

Based on the following data of monthly sales and costs for three companies, perform a regression analysis and calculate the fixed and variable cost ratios for each company. What can you infer from the results? (There is no relationship among the three companies.)

	Company L		Company M		Company N	
	Sales	Costs	Sales	Costs	Sales	Costs
Apr-04	211	187	222	211	208	188
May-04	228	199	228	213	225	200
Jun-04	246	255	246	225	245	235
Jul-04	264	232	264	232	263	235
Aug-04	277	244	277	235	275	241
Sep-04	270	241	270	236	266	241
Oct-04	262	233	262	233	259	233
Nov-04	242	222	242	227	240	222
Dec-04	230	247	230	215	230	220
Jan-05	225	207	235	217	235	217
Feb-05	215	195	225	201	225	208
Mar-05	237	208	247	220	247	220
Apr-05	213	189	247	251	223	208
May-05	230	202	255	254	240	218
Jun-05	248	259	271	265	258	225
Jul-05	268	235	290	271	278	228
Aug-05	277	246	303	273	287	231
Sep-05	271	242	295	278	281	228
Oct-05	264	235	290	273	274	222
Nov-05	244	224	269	267	254	218
Dec-05	234	250	255	253	244	215
Jan-06	227	209	262	260	247	211
Feb-06	216	195	250	239	236	208
Mar-06	239	210	273	261	259	222

Problem 3-7: Anscombe's Quartet

Using Excel, calculate a correlation coefficient and a regression coefficient for each of the following cases. Then draw a scatter plot for each case.

I		II		III		IV	
x	y	x	y	x	y	x	y
10.0	8.04	10.0	9.14	10.0	7.46	8.0	6.58
8.0	6.95	8.0	8.14	8.0	6.77	8.0	5.76
13.0	7.58	13.0	8.74	13.0	12.74	8.0	7.71
9.0	8.81	9.0	8.77	9.0	7.11	8.0	8.84
11.0	8.33	11.0	9.26	11.0	7.81	8.0	8.47
14.0	9.96	14.0	8.1	14.0	8.84	8.0	7.04
6.0	7.24	6.0	6.13	6.0	6.08	8.0	5.25
4.0	4.26	4.0	3.1	4.0	5.39	19.0	12.5
12.0	10.84	12.0	9.13	12.0	8.15	8.0	5.56
7.0	4.82	7.0	7.26	7.0	6.42	8.0	7.91
5.0	5.68	5.0	4.74	5.0	5.73	8.0	6.89

Problem 3-8: Modeling for Estimation

(1) You are planning opening a conveyor belt sushi restaurant which offers all items at a single price in a certain downtown district of the Tokyo Metropolitan area. What model formula can you think of to estimate revenues and expenses? (Assume that you have made sushi before.)

(2) You are considering starting a store which offers coffee at a low price in a relaxed atmosphere. Although it is a highly competitive market, you are confident that your pricing is competitive for the level of quality you offer, and believe you could gain substantial market share. Formulate a model to estimate the demand for low price coffee that people drink at places other than home in terms of the number of cups consumed in a month within the 23 Tokyo Metropolitan Wards.

Problem 3-9: Modeling Convenience Stores

(1) Choose a convenience store chain that you are familiar with and try to analyze and model the most recent sales data for the whole store (not the head office, but total sales for all stores). Collect the data that you need such as sales, number of stores, number of customers, average customer spend etc. from the company's websites or from the websites of industry associations. Please do your utmost to estimate any figures that are unavailable.

(2) What should they do to grow overall sales at the convenience store chain? Consider the model you created in task (1) and investigate the measures that are actually implemented at each chain store in the network, and try to build a model that ties everything together.

END