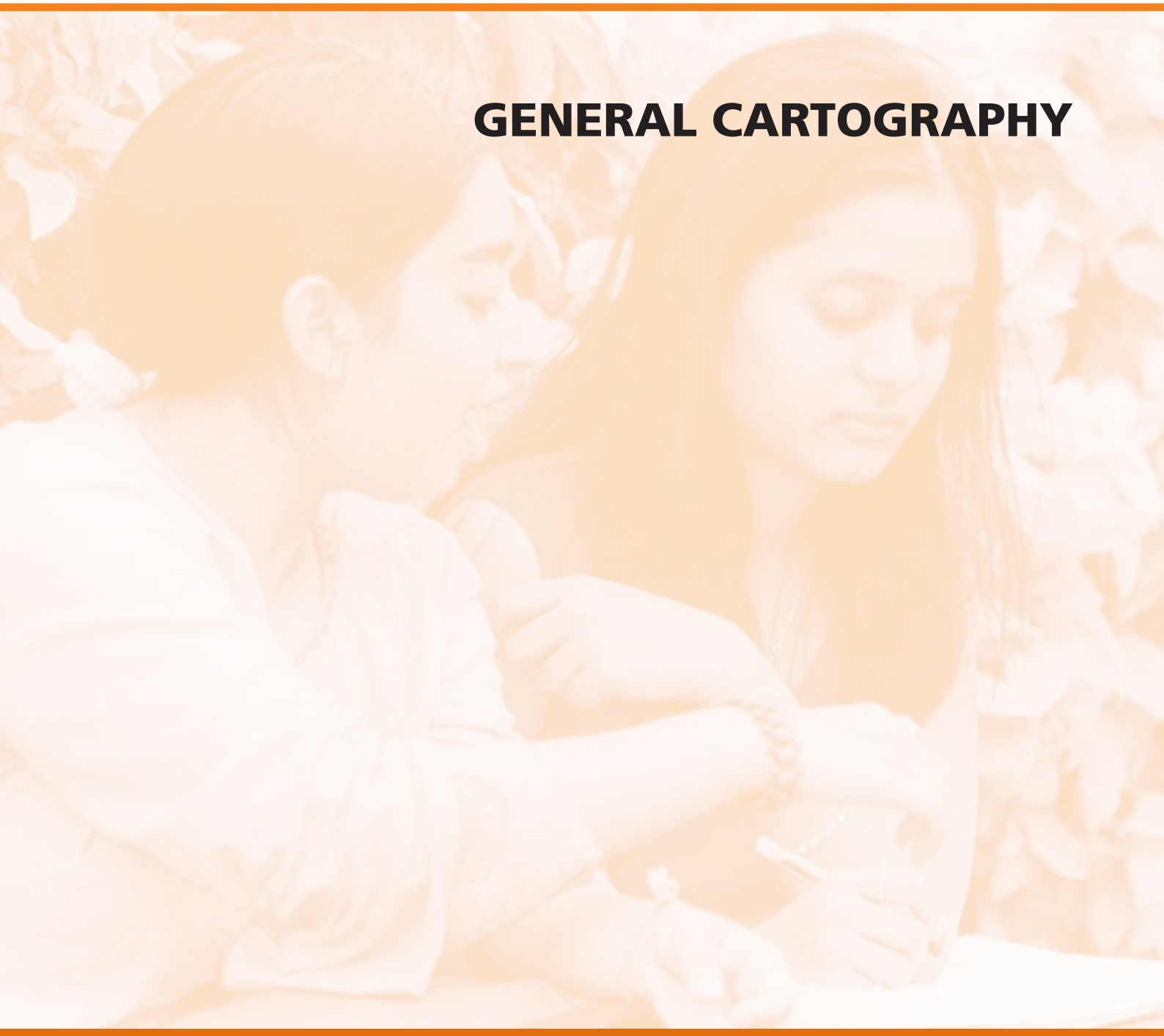
EIILM UNIVERSITY
SIKKIM

GENERAL CARTOGRAPHY



Subject: GENERAL CARTOGRAPHY **Credits: 4**

SYLLABUS

Two Dimensional Diagrams

Divided Rectangle, Wheel Diagram, Square Diagram

Three Dimensional Diagrams

Cube Diagrams, Proportional Spheres, Pictograms

Distributional Map

Isopleth Maps, Choropleth Maps, Dot Maps, Flow-line maps

Statistical Methods

Measurements of central tendencies-Mean, Medium and mode from simple, discrete and continuous series.

Suggested Readings

1. Erwin Raisz, General cartography, McGraw-Hill Book Company
2. George Richard Peter Lawrence, Cartographic methods, Methuen
3. Arthur Howard Robinson, Elements of cartography, Wiley

Chapter 1 - **Two Dimensional Diagrams (Main Topic)**

Learning Objectives

- To define the Cartography.
- To explain the history of Cartography.
- To explain the Diagrams.
- To describe the Square Diagram.

1.1 Cartography

Cartography is the study and practice of making maps. Combining science, aesthetics, and technique, cartography builds on the premise that reality can be modeled in ways that communicate spatial information effectively.

The fundamental problems of traditional cartography are :

- Set the map's agenda and select traits of the object to be mapped. This is the concern of map editing. Traits may be physical, such as roads or land masses, or may be abstract, such as toponyms or political boundaries.
- Represent the terrain of the mapped object on flat media. This is the concern of map projections.
- Eliminate characteristics of the mapped object that are not relevant to the map's purpose. This is the concern of generalization.
- Reduce the complexity of the characteristics that will be mapped. This is also the concern of generalization.
- Orchestrate the elements of the map to best convey its message to its audience. This is the concern of map design.

Modern cartography is largely integrated with geographic information science (GIScience) and constitutes many theoretical and practical foundations of geographic information systems.

1.1.1 **History of cartography**

Cartography, or map-making, has been an integral part of the human history for a long time, possibly up to 8,000 years. From cave paintings to ancient maps of Babylon, Greece, and Asia, through the Age of Exploration, and on into the 21st century, people have created and used maps as essential tools to help them define, explain, and navigate their way through the world. Maps began as two-dimensional drawings but can also adopt three-dimensional shapes (globes, models) and be stored in purely numerical forms.

1.1.1.1 **Earliest known maps**

The earliest known maps are of the heavens, not the earth. Dots dating to 16,500 BC found on the walls of the Lascaux caves map out part of the night sky, including the three bright stars Vega, Deneb, and Altair (the Summer Triangle asterism), as well as the Pleiades star cluster. The Cuevas de El Castillo in Spain contain a dot map of the Corona Borealis constellation dating from 12,000 BC.

Cave painting and rock carvings used simple visual elements that may have aided in recognizing landscape features, such as hills or dwellings. A map-like representation of a mountain, river, valleys and routes around Pavlov in the Czech Republic has been dated to 25,000 BP, and a 14,000 BP polished chunk of sandstone from a cave in Spanish Navarre may represent similar features superimposed on animal etchings, although it may also represent a spiritual landscape, or simply incisings.

Another ancient picture that resembles a map was created in the late 7th millennium BC in Çatalhöyük, Anatolia, modern Turkey. This wall painting may represent a plan of this Neolithic village; however, recent scholarship has questioned the identification of this painting as a map.

Whoever visualized the Çatalhöyük "mental map" may have been encouraged by the fact that houses in Çatalhöyük were clustered together and were entered via flat roofs. Therefore, it was normal for the inhabitants to view their city from a bird's eye view. Later civilizations followed the same convention; today, almost all maps are drawn as if we are looking down from the sky instead of from a horizontal or oblique perspective. The logical advantage of such a perspective is that it provides a view of a greater area, conceptually. There are exceptions: one of the "quasi-maps" of the Minoan civilization on Crete, the “House of the Admiral” wall painting, dating from c. 1600 BC, shows a seaside community in an oblique perspective.

1.1.1.2 **Ancient Near East**

Maps in Ancient Babylonia were made by using accurate surveying techniques.

For example, a 7.6 × 6.8 cm clay tablet found in 1930 at Ga-Sur, near contemporary Kirkuk, shows a map of a river valley between two hills. Cuneiform inscriptions label the features on the map, including a plot of land described as 354 iku (12 hectares) that was owned by a person called Azala. Most scholars date the tablet to the 25th to 24th century BC; Leo Bagrow dissents with a date of 7000 BC. Hills are shown by overlapping semicircles, rivers by lines, and cities by circles. The map also is marked to show the cardinal directions.

An engraved map from the Kassite period (14th–12th centuries BC) of Babylonian history shows the walls and buildings in the holy city of Nippur.

In contrast, the Babylonian World Map, the earliest surviving map of the world (c. 600 BC), is a symbol, not a literal representation. It deliberately omits peoples such as the Persians and Egyptians, who were well known to the Babylonians. The area shown is depicted as a circular shape surrounded by water, which fits the religious image of the world in which the Babylonians believed.

Examples of maps from ancient Egypt are quite rare. However, those that have survived show an emphasis on geometry and well-developed surveying techniques, perhaps stimulated by the need to re-establish the exact boundaries of properties after the annual Nile floods. The Turin Papyrus Map, dated c. 1160 BC, shows the mountains east of the Nile where gold and silver were mined, along with the location of the miners' shelters, wells, and the road network that linked the region with the mainland. Its originality can be seen in the map's inscriptions, its precise orientation, and the use of colour.

1.1.1.3 **Ancient Greece**

1.1.1.3.1 **Early Greek literature**

In reviewing the literature of early geography and early conceptions of the earth, all sources lead to Homer, who is considered by many (Strabo, Kish, and Dilke) as the founding father of Geography. Regardless of the doubts about Homer's existence, one thing is certain: he never was a mapmaker. The enclosed map, which represents the conjectural view of the Homeric world, was never created by him. It is an imaginary reconstruction of the world as Homer described it in his two poems the *Iliad* and the *Odyssey*. It is worth mentioning that each of these writings involves strong geographic symbolism. They can be seen as descriptive pictures of life and warfare in the Bronze Age and the illustrated plans of real journeys. Thus, each one develops a philosophical view of the world, which makes it possible to show this information in the form of a map.

The depiction of the earth conceived by Homer, which was accepted by the early Greeks, represents a circular flat disk surrounded by a constantly moving stream of Ocean (Brown, 22), an idea which would be suggested by the appearance of the horizon as it is seen from a mountaintop or from a seacoast. Homer's knowledge of the Earth was very limited. He and his Greek contemporaries knew very little of the earth beyond Egypt as far south as the Libyan desert, the south-west coast of Asia Minor, and the northern boundary of the Greek homeland. Furthermore, the coast of the Black Sea was only known through myths and legends that circulated during his time. In his poems there is no mention of Europe and Asia as geographical concepts (Thompson, 21), and no mention of the Phoenicians either (Thompson, 40). This seems strange if we recall that the origin of the name Oceanus, a term used by Homer in his poems, belonged to the Phoenicians (Thomson, 27). That is why the big part of Homer's world that is portrayed on this interpretive map represents the lands that border on the Aegean Sea. It is worth noting that even though Greeks believed that they were in the middle of the earth, they also thought that the edges of the world's disk were inhabited by savage, monstrous barbarians and strange animals and monsters; Homer's Odyssey mentions a great many of them.

Additional statements about ancient geography may be found in Hesiod's poems, probably written during the 8th century BC (Kirsh, 1). Through the lyrics of *Works and Days* and *Theogony* he shows to his contemporaries some definite geographical knowledge. He introduces the names of such rivers as Nile, Ister (Danube), the shores of the Bosporus, and the Euxine (Black Sea), the coast of Gaul, the island of Sicily, and a few other regions and rivers (Keane, 6–7). His advanced geographical knowledge not only had predated Greek colonial expansions, but also was used in the earliest Greek world maps, produced by Greek mapmakers such as Anaximander and Hecataeus of Miletus.

1.1.1.3.2 **Early Greek maps**

In classical antiquity, maps were drawn by Anaximander, Hecataeus of Miletus, Herodotus, Eratosthenes, and Ptolemy using both observations by explorers and a mathematical approach.

Early steps in the development of intellectual thought in ancient Greece belonged to Ionians from their well-known city of Miletus in Asia Minor. Miletus was placed favorably to absorb aspects of Babylonian knowledge and to profit from the expanding commerce of the Mediterranean. The earliest ancient Greek who is said to have constructed a map of the world is Anaximander of Miletus (c. 611–546 BC), pupil of Thales. He believed that the earth was a cylindrical form, like a stone pillar and suspended in space. The inhabited part of his world was circular, disk-shaped, and presumably located on the upper surface of the cylinder (Brown, 24).

Anaximander was the first ancient Greek to draw a map of the known world. It is for this reason that he is considered by many to be the first mapmaker (Dilke, 23). A scarcity of archaeological and written evidence prevents us from giving any assessment of his map. What we may presume is that he portrayed land and sea in a map form. Unfortunately, any definite geographical knowledge that he included in his map is lost as well. Although the map has not survived, Hecataeus of Miletus (550–475 BC) produced another map fifty years later that he claimed was an improved version of the map of his illustrious predecessor.

Hecatæus's map describes the earth as a circular plate with an encircling Ocean and Greece in the center of the world. This was a very popular contemporary Greek worldview, derived originally from the Homeric poems. Also, similar to many other early maps in antiquity his map has no scale. As units of measurements, this map used "days of sailing" on the sea and "days of marching" on dry land (Goode, 2). The purpose of this map was to accompany Hecatæus's geographical work that was called *Periodos Ges*, or *Journey Round the World* (Dilke, 24). *Periodos Ges* was divided into two books, "Europe" and "Asia", with the latter including Libya, the name of which was an ancient term for all of the known Africa.

The work follows the assumption of the author that the world was divided into two continents, Asia and Europe. He depicts the line between the Pillars of Hercules through the Bosporus, and the Don River as a boundary between the two. Hecatæus is the first known writer who thought that the Caspian flows into the circumference ocean—an idea that persisted long into the Hellenic period. He was particularly informative on the Black Sea, adding many geographic places that already were known to Greeks through the colonization process. To the north of the Danube, according to Hecatæus, were the Rhipæan (gusty) Mountains, beyond which lived the Hyperboreans—peoples of the far north. Hecatæus depicted the origin of the Nile River at the southern circumference ocean. His view of the Nile seems to have been that it came from the southern circumference ocean. This assumption helped Hecatæus solve the mystery of the annual flooding of the Nile. He believed that the waves of the ocean were a primary cause of this occurrence (Tozer, 63). It is worth mentioning that a similar map based upon one designed by Hecataeus was intended to aid political decision-making. According to Herodotus, it was engraved upon a bronze tablet and was carried to Sparta by Aristagoras during the revolt of the Ionian cities against Persian rule from 499 to 494 BC.

Anaximenes of Miletus (6th century BC), who studied under Anaximander, rejected the views of his teacher regarding the shape of the earth and instead, he visualized the earth as a rectangular form supported by compressed air.

Pythagoras of Samos (c. 560–480 BC) speculated about the notion of a spherical earth with a central fire at its core. He is also credited with the introduction of a model that divides a spherical earth into five zones: one hot, two temperate, and two cold—northern and southern. It seems likely that he illustrated his division in the form of a map, however, no evidence of this has survived to the present.

Scylax, a sailor, made a record of his Mediterranean voyages inc. 515 BC. This is the earliest known set of Greek periploi, or sailing instructions, which became the basis for many future mapmakers, especially in the medieval period.

The way in which the geographical knowledge of the Greeks advanced from the previous assumptions of the Earth's shape was through Herodotus and his conceptual view of the world. This map also did not survive and many have speculated that it was never produced.

Herodotus traveled very extensively, collecting information and documenting his findings in his books on Europe, Asia, and Libya. He also combined his knowledge with what he learned from the people he met.

Herodotus wrote his *Histories* in the mid-5th century BC. Although his work was dedicated to the story of the long struggle of the Greeks with the Persian Empire, Herodotus also included everything he knew about the geography, history, and peoples of the world. Thus, his work provides a detailed picture of the known world of the 5th century BC.

Herodotus rejected the prevailing view of most 5th century BC maps that the earth is a circular plate surrounded by Ocean. In his work he describes the earth as an irregular shape with oceans surrounding only Asia and Africa. He introduces names such as the Atlantic Sea and the Erythrean Sea. He also divided the world into three continents: Europe, Asia, and Africa. He depicted the boundary of Europe as the line from the Pillars of Hercules through the Bosporus and the area between the Caspian Sea and Indus River. He regarded the Nile as the boundary between Asia and Africa. He speculated that the extent of Europe was much greater than was assumed at the time and left Europe's shape to be determined by future research.

In the case of Africa, he believed that, except for the small stretch of land in the vicinity of Suez, the continent was in fact surrounded by water. However, he definitely disagreed with his predecessors and contemporaries about its presumed circular shape. He based his theory on the story of Pharaoh Necho II, the ruler of Egypt between 609 and 594 BC, who had sent Phoenicians to circumnavigate Africa. Apparently, it took them three years, but they certainly did prove his idea. He speculated that the Nile River started as far west as the Ister River in Europe and cut Africa through the middle. He was the first writer to assume that the Caspian Sea was separated from other seas and he recognized northern Scythia as one of the coldest inhabited lands in the world.

Similar to his predecessors, Herodotus also made mistakes. He accepted a clear distinction between the civilized Greeks in the center of the earth and the barbarians on the world's edges. In his *Histories* we can see very clearly that he believed that the world became stranger and stranger when one traveled away from Greece, until one reached the ends of the earth, where humans behaved as savages.

1.1.1.3.3 **Spherical Earth and meridians**

Whereas a number of previous Greek philosophers presumed the earth to be spherical, Aristotle (384–322 BC) is the one to be credited with proving the Earth's sphericity. Those arguments may be summarized as follows:

- The lunar eclipse is always circular
- Ships seem to sink as they move away from the view and pass the horizon
- Some stars can be seen only from certain parts of the Earth.

A vital contribution to mapping the reality of the world came with a scientific estimate of the circumference of the earth. This event has been described as the first scientific attempt to give geographical studies a mathematical basis. The man credited for this achievement was Eratosthenes (275–195 BC). As described by George Sarton, historian of science, "there was among them [Eratosthenes's contemporaries] a man of genius but as he was working in a new field they were too stupid to recognize him" (Noble, 27). His work, including *On the Measurement of the Earth* and *Geographica*, has only survived in the writings of later philosophers such as Cleomedes and Strabo. He was a devoted geographer who sets out to reform and perfect the map of the world. Eratosthenes argued that accurate mapping, even if in two dimensions only, depends upon the establishment of accurate linear measurements. He was the first to calculate the circumference of the Earth (within 0.5 percent accuracy) by calculating the heights of shadows on different parts of the Egypt at a given time. The first in Alexandria, the other further up the Nile, in the Ancient Egyptian city of Swenet (known in Greek as

Syene) where reports of a well into which the sun shone only on the summer solstice, long existed. Proximity to the Tropic of Cancer being the dynamics creating the effect. He had the distance between the two shadows calculated and then their height. From this he determined the difference in angle between the two points and calculated how large a circle would be made by adding the rest of the degrees to 360. His great achievement in the field of map-making was the use of a new technique of charting with meridians, his imaginary north–south lines, and parallels, his imaginary west–east lines. These axis lines were placed over the map of the earth with their origin in the city of Rhodes and divided the world into sectors. Then, Eratosthenes used these earth partitions to reference places on the map. He also was the first person to divide Earth correctly into five climatic regions: a torrid zone across the middle, two frigid zones at the extreme north and south, and two temperate bands in between. He was also the first person to use the word "geography".

Claudius Ptolemy (90–168) thought that, with the aid of astronomy and mathematics, the earth could be mapped very accurately. Ptolemy revolutionized the depiction of the spherical earth on a map by using perspective projection, and suggested precise methods for fixing the position of geographic features on its surface using a coordinate system with parallels of latitude and meridians of longitude.

Ptolemy's eight-volume atlas *Geographia* is a prototype of modern mapping and GIS. It included an index of place-names, with the latitude and longitude of each place to guide the search, scale, conventional signs with legends, and the practice of orienting maps so that north is at the top and east to the right of the map—an almost universal custom today.

Yet with all his important innovations, however, Ptolemy was not infallible. His most important error was a miscalculation of the circumference of the earth. He believed that Eurasia covered 180° of the globe, which convinced Christopher Columbus to sail across the Atlantic to look for a simpler and faster way to travel to India. Had Columbus known that the true figure was much greater, it is conceivable that he would never have set out on his momentous voyage.

1.1.1.4 **Roman Empire**

1.1.1.4.1 **Pomponius Mela**

Pomponius is unique among ancient geographers in that, after dividing the earth into five zones, of which two had only been habitable, he asserts the existence of antichthones, inhabiting the southern temperate zone inaccessible to the folk of the northern temperate regions from the unbearable heat of the intervening torrid belt. On the divisions and boundaries of Europe, Asia and Africa, he repeats Eratosthenes; like all classical geographers from Alexander the Great (except Ptolemy) he regards the Caspian Sea as an inlet of the Northern Ocean, corresponding to the Persian Gulf and the Red Sea on the south.

1.1.1.4.2 **5th-century Roman road map**

In 2007, the Tabula Peutingeriana, a 12th-century replica of a 5th-century map, was placed on the UNESCO Memory of the World Register and displayed to the public for the first time. Although well preserved and believed to be an accurate copy of an authentic original, the scroll media it is on is so delicate now it must be protected at all times from exposure to daylight.

1.1.1.5 **China**

1.1.1.5.1 **Earliest extant maps from the Qin State**

The earliest maps known to have survived in China date to the 4th century BC. In 1986, seven ancient Chinese maps were found in an archeological excavation of a Qin State tomb in what is now Fangmatan, in the vicinity of Tianshui City, Gansu province. Before this find, the earliest extant maps that were known came from the Mawangdui excavation in 1973, which found three maps on silk dated to the 2nd century BC in the early Han Dynasty. The 4th century BCE maps of the State of Qin were drawn with black ink on wooden blocks. These blocks fortunately survived in soaking conditions due to underground water that had seeped into the tomb; the quality of the wood had much to do with their survival. After two years of slow-drying techniques, the maps were fully restored.

The territory shown in the seven Qin maps overlaps each other. The maps display tributary river systems of the Jialing River in Sichuan province, in a total measured area of 107 by 68 km. The maps featured rectangular symbols encasing character names for the locations of administrative counties. Rivers and roads are displayed with similar line symbols; this makes interpreting the map somewhat difficult, although the labels of rivers placed in order on stream flow are helpful to modern day cartographers. These maps also feature locations where different types of timber can be gathered, while two of the maps states the distances in mileage to the timber sites. In light of this, these maps are perhaps the oldest economic maps in the world since they predate Strabo's economic maps.

In addition to the seven maps on wooden blocks found at Tomb 1 of Fangmatan, a fragment of a paper map was found on the chest of the occupant of Tomb 5 of Fangmatan in 1986. This tomb is dated to the early Western Han, so the map dates to the early 2nd century BC. The map shows topographic features such as mountains, waterways and roads, and is thought to cover the area of the preceding Qin Kingdom.

1.1.1.5.2 **Earliest geographical writing**

In China, the earliest known geographical Chinese writing dates back to the 5th century BC, during the beginning of the Warring States (481–221 BC). This was the 'Yu Gong' ('Tribute of Yu') chapter of the book *Shu Jing* (*Classic of History*). The book describes the traditional nine provinces, their kinds of soil, their characteristic products and economic goods, their tributary goods, their trades and vocations, their state revenues and agricultural systems, and the various rivers and lakes listed and placed accordingly. The nine provinces in the time of this geographical work were very small in terrain size compared to what modern China occupies today. In fact, its description pertained to areas of the Yellow River, the lower valleys of the Yangtze, with the plain between them and the Shandong Peninsula, and to the west the most northern parts of the Wei River and the Han River were known (along with the southern parts of modern day Shanxi province).

1.1.1.5.3 **Earliest known reference to a map, or 'tu'**

The oldest reference to a map in China comes from the 3rd century BC. This was the event of 227 BC where Crown Prince Dan of Yan had his assassin Jing Ke visit the court of the ruler of the State of Qin, who would become Qin Shi Huang (r. 221–210 BC). Jing Ke was to present the ruler of Qin with a district map painted on a silk scroll, rolled up and held in a case where he hid his assassin's dagger. Handing him the map of the designated territory was the first diplomatic act of submitting that district to Qin rule. Instead he attempted to kill Qin, an assassination plot that failed. From then on maps are frequently mentioned in Chinese sources.

1.1.1.5.4 **Han Dynasty and period of division**

The three Han Dynasty maps found at Mawangdui differ from the earlier Qin State maps. While the Qin maps place the cardinal direction of north at the top of the map, the Han maps are orientated with the southern direction at the top. The Han maps are also more complex, since they cover a much larger area, employ a large number of well-designed map symbols, and include additional information on local military sites and the local population. The Han maps also note measured distances between certain places, but a formal graduated scale and rectangular grid system for maps would not be used—or at least described in full—until the 3rd century (see Pei Xiu below). Among the three maps found at Mawangdui was a small map representing the tomb area where it was found, a larger topographical map showing the Han's borders along the subordinate Kingdom in Changsha and the Nanyue Kingdom (of northern Vietnam and parts of modern Guangdong and Guangxi), and a map which marks the positions of Han military garrisons that were employed in an attack against Nanyue in 181 BC.

An early text that mentioned maps was the *Rites of Zhou*. Although attributed to the era of the Zhou Dynasty, its first recorded appearance was in the libraries of Prince Liu De (c. 130 BC), and was compiled and commented on by Liu Xin in the 1st century AD. It outlined the use of maps that were made for governmental provinces and districts, principalities, frontier boundaries, and even pinpointed locations of ores and minerals for mining facilities. Upon the investiture of three of his sons as feudal princes in 117 BC, Emperor Wu of Han had maps of the entire empire submitted to him.

From the 1st century AD onwards, official Chinese historical texts contained a geographical section (Diliji), which was often an enormous compilation of changes in place-names and local administrative divisions controlled by the ruling dynasty, descriptions of mountain ranges, river systems, taxable products, etc. From the time of the 5th century BC *Shu Jing* forward, Chinese geographical writing provided more concrete information and less legendary element. This example can be seen in the 4th chapter of the *Huainanzi* (Book of the Master of Huainan), compiled under the editorship of Prince Liu An in 139 BC during the Han Dynasty (202 BC–202 AD). The chapter gave general descriptions of topography in a systematic fashion, given visual aids by the use of maps (di tu) due to the efforts of Liu An and his associate Zuo Wu. In Chang Chu's *Hua Yang Guo Chi* (*Historical Geography of Szechuan*) of 347, not only rivers, trade routes, and various tribes were described, but it also wrote of a 'Ba June Tu Jing' ('Map of Szechuan'), which had been made much earlier in 150.

Local map-making such as the one of Szechuan mentioned above, became a widespread tradition of Chinese geographical works by the 6th century, as noted in the bibliography of the *Sui Shu*. It is during this time of the Southern and Northern Dynasties that the Liang Dynasty (502–557) cartographers also began carving maps into stone steles (alongside the maps already drawn and painted on paper and silk).

1.1.1.5.5 **Pei Xiu, the 'Ptolemy of China'**

In the year 267, Pei Xiu (224–271) was appointed as the Minister of Works by Emperor Wu of Jin, the first emperor of the Jin Dynasty. Pei is best known for his work in map-making. Although map making and use of the grid existed in China before him, he was the first to mention a plotted geometrical grid and graduated scale displayed on the surface of maps to gain greater accuracy in the estimated distance between different locations. Pei outlined six principles that should be observed when creating maps, two of which included the rectangular grid and the graduated scale for measuring distance. Historians compare him to the Greek Ptolemy for his contributions in map-making. However, Howard Nelson states that, although the accounts of earlier cartographic works by the inventor and official Zhang Heng (78–139) are somewhat vague and sketchy, there is ample written evidence that Pei Xiu derived the use of the rectangular grid reference from the maps of Zhang Heng.

Later Chinese ideas about the quality of maps made during the Han Dynasty and before stem from the assessment given by Pei Xiu, which was not a positive one. Pei Xiu noted that the extant Han maps at his disposal were of little use since they featured too many inaccuracies and exaggerations in measured distance between locations. However, the Qin State maps and Mawangdui maps of the Han era were far superior in quality than those examined by Pei Xiu. It was not until the 20th century that Pei Xiu's 3rd century assessment of earlier maps' dismal quality would be overturned and disproven. The Qin and Han maps did have a degree of accuracy in scale and pinpointed location, but the major improvement in Pei Xiu's work and that of his contemporaries was expressing topographical elevation on maps.

1.1.1.5.6 **Sui and Tang dynasties**

In the year 605, during the Sui Dynasty (581–618), the Commercial Commissioner Pei Ju (547–627) created a famous geometrically gridded map. In 610 Emperor Yang of Sui ordered government officials from throughout the empire to document in gazetteers the customs, products, and geographical features of their local areas and provinces, providing descriptive writing and drawing them all onto separate maps, which would be sent to the imperial secretariat in the capital city.

The Tang Dynasty (618–907) also had its fair share of cartographers, including the works of Xu Jingzong in 658, Wang Mingyuan in 661, and Wang Zhongsi in 747. Arguably the greatest geographer and cartographer of the Tang period was Jia Dan (730–805), whom Emperor Dezong of Tang entrusted in 785 to complete a map of China with her recently former inland colonies of Central Asia, the massive and detailed work completed in 801, called the *Hai Nei Hua Yi Tu* (Map of both Chinese and Barbarian Peoples within the (Four) Seas). The map was 30 ft long (9.1 m) and 33 ft high (10 m) in dimension, mapped out on a grid scale of 1-inch (25 mm) equaling 100 li (unit) (the Chinese equivalent of the mile/kilometer). Jia Dan is also known for having described the Persian Gulf region with great detail, along with lighthouses that were erected at the mouth of the Persian Gulf by the medieval Iranians during the Abbasid period (refer to the article on Tang Dynasty for more).

1.1.1.5.7 **Song Dynasty**

During the Song Dynasty (960–1279) Emperor Taizu of Song ordered Lu Duosun in 971 to update and 're-write all the Tu Jing in the world', which would seem to be a daunting task for one individual, who was sent out throughout the provinces to collect texts and as much data as possible. With the aid of Song Zhun, the massive work was completed in 1010, with some 1566 chapters. The later *Song Shi* historical text stated (Wade-Giles spelling):

> “ Yuan Hsieh (d. +1220) was Director-General of governmental grain stores. In pursuance of his schemes for the relief of famines he issued orders that each pao (village) should prepare a map which would show the fields and mountains, the rivers and the roads in fullest detail. The maps of all the pao were joined together to make a map of the tu (larger district), and these in turn were joined with others to make a map of the hsiang and the hsien (still larger districts). If there was any trouble about the collection of taxes or the distribution of grain, or if the question of chasing robbers and bandits arose, the provincial officials could readily carry out their duties by the aid of the maps.

Like the earlier Liang Dynasty stone-stele maps (mentioned above), there were large and intricately carved stone stele maps of the Song period. For example, the 3 ft (0.91 m) squared stone stele map of an anonymous artist in 1137, following the grid scale of 100 li squared for each grid square. What is truly remarkable about this map is the incredibly precise detail of coastal outlines and river systems in China (refer to Needham's Volume 3, Plate LXXXI for an image). The map shows 500 settlements and a dozen

rivers in China, and extends as far as Korea and India. On the reverse, a copy of a more ancient map uses grid coordinates in a scale of 1:1,500,000 and shows the coastline of China with great accuracy.

The famous 11th century scientist and polymath statesman Shen Kuo (1031–1095) was also a geographer and cartographer. His largest atlas included twenty three maps of China and foreign regions that were drawn at a uniform scale of 1:900,000. Shen also created a three-dimensional raised-relief map using sawdust, wood, beeswax, and wheat paste, while representing the topography and specific locations of a frontier region to the imperial court. Shen Kuo's contemporary, Su Song (1020–1101), was a cartographer who created detailed maps in order to resolve a territorial border dispute between the Song Dynasty and the Liao Dynasty.

1.1.1.5.8 **Ming and Qing dynasties**

The Da Ming hunyi tu map, dating from about 1390, is in multicolor. The horizontal scale is 1:820,000 and the vertical scale is 1:1,060,000.

In 1579, Luo Hongxian published the Guang Yutu atlas, including more than 40 maps, a grid system, and a systematic way of representing major landmarks such as mountains, rivers, roads and borders. The Guang Yutu incorporates the discoveries of naval explorer Zheng He's 15th century voyages along the coasts of China, Southeast Asia, India and Africa.

From the 16th and 17th centuries, several examples survive of maps focused on cultural information. Gridlines are not used on either Yu Shi's Gujin xingsheng zhi tu (1555) or Zhang Huang's Tushu bian (1613); instead, illustrations and annotations show mythical places, exotic foreign peoples, administrative changes and the deeds of historic and legendary heroes. Also in the 17th century, an edition of a possible Tang Dynasty map shows clear topographical contour lines. Although topographic features were part of maps in China for centuries, a Fujian county official Ye Chunji (1532–1595) was the first to base county maps using on-site topographical surveying and observations.

The Korean made Kangnido based on two Chinese maps, which describes the Old World.

1.1.1.5.9 **Mongol Empire**

In the Mongol Empire, the Mongol scholars with the Persian and Chinese cartographers or their foreign colleagues created maps, geographical compendium as well as travel accounts. Rashid-al-Din Hamadani described his geographical compendium, "Suvar al-aqalim", constituted volume four of the Collected chronicles of the Ilkhanate in Persia. His works say about the borders of the seven climes (old world), rivers, major cities, places, climate, and Mongol yams (relay stations). The Great Khan Khubilai's ambassador and minister, Bolad, had helped Rashid's works in relation to the Mongols and Mongolia. Thanks to Pax Mongolica, the easterners and the westerners in Mongol dominions were able to gain access to one another's geographical material.

The Mongols required the nations they conquered to send geographical maps to the Mongol headquarters.

One of medieval Persian work written in northwest Iran can clarify the historical geography of Mongolia where Genghis Khan was born and united the Mongol and Turkic nomads as recorded in native sources, especially the Secret History of the Mongols.

Map of relay stations, called "yam", and strategic points existed in the Yuan Dynasty. The Mongol map-making was enriched by traditions of ancient China and Iran which were now under the Mongols.

Because the Yuan court often requested the western Mongol khanates to send their maps, the Yuan Dynasty was able to publish a map describing the whole Mongol world in c.1330. This is called "Hsi-pei pi ti-li tu". The map includes the Mongol dominions including 30 cities in Iran such as Ispahan and the Ilkhanid capital Soltaniyeh, and Russia (as "Orash") as well as their neighbors, e.g. Egypt and Syria.

1.1.1.6 Cartography of India

The **cartography of India** begins with early charts for navigation and constructional plans for buildings. Indian traditions influenced Tibetan and Islamic traditions, and in turn, were influenced by the British cartographers who solidified modern concepts into India's map making.

A prominent foreign geographer and cartographer was Hellenistic geographer Ptolemy (90–168) who researched at the library in Alexandria to produce a detailed eight-volume record of world geography. During the Middle Ages, India sees some exploration by Chinese and Muslim geographers, while European maps of India remain very sketchy. A prominent medieval cartographer was Persian geographer Abu Rayhan Biruni (973–1048) who visited India and studied the country's geography extensively.

European maps become more accurate with the Age of Exploration and Portuguese India from the 16th century. The first modern maps were produced by Survey of India, established in 1767 by the British East India Company. Survey of India remains in continued existence as the official mapping authority of the Republic of India.

1.1.1.6.1 **Prehistory**

Joseph E. Schwartzberg (2008) proposes that the Bronze Age Indus Valley Civilization (c. 2500–1900 BCE) may have known "cartographic activity" based on a number of excavated surveying instruments and measuring rods and that the use of large scale constructional plans, cosmological drawings, and cartographic material was known in India with some regularity since the Vedic period (1st millennium BCE).

> 'Though not numerous, a number of map-like graffiti appear among the thousands of Stone Age Indian cave paintings; and at least one complex Mesolithic diagram is believed to be a representation of the cosmos.'

Susan Gole (1990) comments on the cartographic traditions in early India:

The fact that towns as far apart as Mohenjodaro near the Indus and Lothal on the Saurashtra coast were built in the second millennium BCE with baked bricks of identical size on similar plans denotes a widespread recognition of the need for accuracy in planning and management. In the 8th century CE the Kailas temple at Ellora in Maharashtra was carved down into mountain for 100 feet, with intricate sculptures lining pillared halls, no easy task even with an exact map to follow, impossible without. So if no maps have been found, it should not be assumed that the Indians did not know how to conceptualize in a cartographic manner.

1.1.1.6.2 **Antiquity**

Map-making of India as a part of the greater continent of Asia develops in Classical Antiquity.

In Greek map-making, India appears as a remote land on the eastern fringe of Asia in the 5th century BCE (Hecataeus of Miletus). More detailed knowledge becomes available after the conquests of Alexander the Great, and the 3rd-century BCE geographer Eratosthenes has a clearer idea of the size and location of India. By the 1st century, at least the western coast of India is well known to Hellenistic geography, with itineraries such as the Periplus of the Erythraean Sea. Ptolemy by the 2nd century has good knowledge of the Indian Sea, including an oversized Sri Lanka (*Taprobane*), but not of the interior of the subcontinent.

Native Indian cartographic traditions before the Hellenistic period remain rudimentary. Early forms of map-making in India included legendary paintings; maps of locations described in Indian epic poetry, for example the *Ramayana*. These works contained descriptions of legendary places, and often even described the nature of the mythological inhabitants of a particular location. Early Indian map-making showed little knowledge of scale, the important parts of the map were shown to be larger than others (Gael 1990). Indian cartographic traditions also covered the locations of the Pole star, and other constellations of use. These charts may have been in use by the beginning of the Common Era for purposes of navigation. Other early maps in India include the Udayagiri wall sculpture—made under the Gupta empire in 400 CE—showing the meeting of the Ganges and the Yamuna.

1.1.1.6.3 **Middle Ages**

The 8th-century scholar Bhavabhuti conceived paintings which indicated geographical regions. The boundaries of land, granted to the Brahman priests of India by their patrons, were described in detail. The descriptions indicated good geographical knowledge and in one case over 75 details of the land granted have been found. The Chinese records of the Tang dynasty show that a map of the neighboring Indian region was gifted to Wang Hiuen-tse by its king.

In the 9th century, Islamic geographers under Abbasid Caliph Al-Ma'mun improved on Ptolemy's work and depicted the Indian Ocean as an open body of water instead of a landlocked sea as Ptolemy had done. The Iranian geographers Abū Muhammad al-Hasan al-Hamdānī and Habash al-Hasib al-Marwazi set the Prime Meridian of their maps at Ujjain, a center of Indian astronomy. In the early 11th century, the Persian geographer Abu Rayhan Biruni visited India and studied the country's geography extensively. He was considered the most skilled when it came to mapping cities and measuring the distances between them, which he did for many cities in the western Indian subcontinent. He also wrote extensively on the geology of India. In 1154, the Arab geographer Muhammad al-Idrisi included a section on the map-making and geography of India and its neighboring countries in his world atlas, *Tabula Rogeriana*.

European scholar Francesco I reproduced a number of Indian maps in his magnum opus *La Cartografia Antica dell India*. Out these maps two have been reproduced using a manuscript of *Lokaprakasa*—originally compiled by the polymath Ksemendra (Kashmir, 11th century CE) —as a source. The other manuscript, used as a source by Francesco I, is titled *Samgrahani*. The early volumes of the *Encyclopædia Britannica* also described cartographic charts made by the Dravidian people of India.

The cartographic tradition of India influenced the map making tradition of Tibet, where maps of Indian origin have been discovered. Islamic map-making was also influenced by the Indian tradition as a result of extensive contact.

1.1.1.6.4 **Mughal era**

Maps from the Ain-e-Akbari, a Mughal document detailing India's history and traditions, contain references to locations indicated in earlier Indian cartographic traditions.

The seamless globe was invented in Kashmir by Ali Kashmiri ibn Luqman in 998 AH (1589-90 CE), and twenty other such globes were later produced in Lahore and Kashmir during the Mughal Empire. Before they were rediscovered in the 1980s, it was believed by modern metallurgists to be technically impossible to produce metal globes without any seams, even with modern technology. These Mughal metallurgists pioneered the method of lost-wax casting in order to produce these globes.

The scholar Sadiq Isfahani of Jaunpur compiled an atlas of the parts of the world which he held to be 'suitable for human life'. The 32 sheet atlas—with maps oriented towards the south as was the case with Islamic works of the era—is part of a larger scholarly work compiled by Isfahani during 1647 CE. According to Joseph E. Schwartzberg (2008): 'The largest known Indian map, depicting the former Rajput capital at Amber in remarkable house-by-house detail, measures 661 × 645 cm. (260 × 254 in., or approximately 22 × 21 ft).'

1.1.1.6.5 **Colonial India**

A map describing the kingdom of Nepal, four feet in length and about two and a half feet in breadth, was presented to Warren Hastings. In this raised-relief map the mountains were elevated above the surface and several geographical elements were indicated in different colors. The Europeans used 'scale-bars' in their cartographic tradition. Upon their arrival in India during the middle ages, the indigenous Indian measures were reported back to Europe, and first published by Guillaume de I'Isle in 1722 as *Carte des Costes de Malabar et de Coromandel*.

With the establishment of the British Raj in India, modern European cartographic traditions were officially employed by the British Survey of India (1767). One British observer commented on the tradition of native Indian map-making.

Besides geographical tracts, the Hindus have also mapped the world according to the system of the puranics and of the astronomers: the latter is very common. They also have maps of India and of particular districts, in which latitudes and longitudes are entirely out of the question, and they never make use of scale of equal parts. The sea shores, rivers and ranges of mountains are represented by straight lines.

The Great Trigonometric Survey, a project of the Survey of India throughout most of the 19th century, was piloted in its initial stages by William Lambton, and later by George Everest. To achieve the highest accuracy a number of corrections were applied to all distances calculated from simple trigonometry:

- Curvature of the earth
- The non spherical nature of the curvature of the earth
- Gravitational influence of mountains on pendulums
- Refraction
- Height above sea level

Thomas George Montgomerie organized several cartographic expeditions to map Tibet, as well as China. Mohamed-i-Hameed, Nain Singh and Mani Singh were among the agents employed by the British for their cartographic operations. Nain Singh, in particular, became famous for his geographical knowledge of Asia, and was awarded several honors for his expeditions.

Modern India (1947 to present)

The modern map making techniques in India, like other parts of the world, employ digitization, photographic surveys and printing. Satellite imageries, aerial photographs and video surveying techniques are also used. The Indian IRS-P5 (CARTOSAT-1) was equipped with high resolution panchromatic equipment to enable it for cartographic purposes. IRS-P5 (CARTOSAT-1) was followed by a more advanced model named IRS-P6 developed also for agricultural applications. The CARTOSAT-2 project, equipped with a single panchromatic camera which supported scene specific on-spot images, succeed the CARTOSAT-1 project.

1.1.1.7 **Arab cartography**

In the Middle Ages, Muslim scholars continued and advanced on the map-making traditions of earlier cultures. Most used Ptolemy's methods; but they also took advantage of what explorers and merchants learned in their travels across the Muslim world, from Spain to India to Africa, and beyond in trade relationships with China, and Russia.

An important influence in the development of map-making was the patronage of the Abbasid caliph, al-Ma'mun, who reigned from 813 to 833. He commissioned several geographers to remeasure the distance on earth that corresponds to one degree of celestial meridian. Thus his patronage resulted in the refinement of the definition of the mile used by Arabs (*mīl* in Arabic) in comparison to the *stadion* used by the Greeks. These efforts also enabled Muslims to calculate the circumference of the earth. Al-Mamun also commanded the production of a large map of the world, which has not survived, though it is known that its map projection type was based on Marinus of Tyre rather than Ptolemy.

Also in the 9th century, the Persian mathematician and geographer, Habash al-Hasib al-Marwazi, employed the use spherical trigonometry and map projection methods in order to convert polar coordinates to a different coordinate system centered on a specific point on the sphere, in this the Qibla, the direction to Mecca. Abū Rayhān Bīrūnī (973–1048) later developed ideas which are seen as an anticipation of the polar coordinate system. Around 1025, he describes a polar equi-azimuthal equidistant projection of the celestial sphere. However, this type of projection had been used in ancient Egyptian star-maps and was not to be fully developed until the 15 and 16th centuries.

In the early 10th century, Abū Zayd al-Balkhī, originally from Balkh, founded the "Balkhī school" of terrestrial mapping in Baghdad. The geographers of this school also wrote extensively of the peoples, products, and customs of the areas in the Muslim world, with little interest in the non-Muslim realms. The "Balkhī school", which included geographers such as Estakhri, al-Muqaddasi and Ibn Hawqal, produced world atlases, each one featuring a world map and twenty regional maps.

Suhrāb, a late 10th-century Muslim geographer, accompanied a book of geographical coordinates with instructions for making a rectangular world map, with equirectangular projection or cylindrical equidistant projection. The earliest surviving rectangular coordinate map is dated to the 13th century and is attributed to Hamdallah al-Mustaqfi al-Qazwini, who based it on the work of Suhrāb. The orthogonal parallel lines were separated by one degree intervals, and the map was limited to Southwest Asia and Central Asia. The earliest surviving world maps based on a rectangular coordinate grid are attributed to al-Mustawfi in the 14th or 15th century (who used intervals of ten degrees for the lines), and to Hafiz-I Abru (died 1430).

Ibn Battuta (1304–1368?) wrote "Rihlah" (Travels) based on three decades of journeys, covering more than 120,000 km through northern Africa, southern Europe, and much of Asia.

1.1.1.7.1 **Regional cartography**

Islamic regional map-making is usually categorized into three groups: that produced by the "Balkhī school", the type devised by Muhammad al-Idrisi, and the type that is uniquely found in the *Book of curiosities*.

The maps by the Balkhī schools were defined by political, not longitudinal boundaries and covered only the Muslim world. In these maps the distances between various "stops" (cities or rivers) were equalized. The only shapes used in the designs were verticals, horizontals, 90-degree angles, and arcs of circles; unnecessary geographical details were eliminated. This approach is similar to that used in subway maps, most notable used in the "London Underground Tube Map" in 1931 by Harry Beck.

Al-Idrīsī defined his maps differently. He considered the extent of the known world to be 160° in longitude, and divided the region into ten parts, each 16° wide. In terms of latitude, he portioned the known world into seven 'climes', determined by the length of the longest day. In his maps, many dominant geographical features can be found.

1.1.1.7.2 ***Book on the appearance of the Earth***

Muhammad ibn Mūsā al-Khwārizmī's *Kitāb ṣūrat al-Arḍ* ("Book on the appearance of the Earth") was completed in 833. It is a revised and completed version of Ptolemy's *Geography*, consisting of a list of 2402 coordinates of cities and other geographical features following a general introduction.

Al-Khwārizmī, Al-Ma'mun's most famous geographer, corrected Ptolemy's gross overestimate for the length of the Mediterranean Sea (from the Canary Islands to the eastern shores of the Mediterranean); Ptolemy overestimated it at 63 degrees of longitude, while al-Khwarizmi almost correctly estimated it at nearly 50 degrees of longitude. Al-Ma'mun's geographers "also depicted the Atlantic and Indian Oceans as open bodies of water, not landlocked seas as Ptolemy had done. " Al-Khwarizmi thus set the Prime Meridian of the Old World at the eastern shore of the Mediterranean, 10–13 degrees to the east of Alexandria (the prime meridian previously set by Ptolemy) and 70 degrees to the west of Baghdad. Most medieval Muslim geographers continued to use al-Khwarizmi's prime meridian. Other prime meridians used were set by Abū Muhammad al-Hasan al-Hamdānī and Habash al-Hasib al-Marwazi at Ujjain, a center of Indian astronomy, and by another anonymous writer in Basra.

1.1.1.7.3 ***Tabula Rogeriana***

The Arab geographer, Muhammad al-Idrisi, produced his medieval atlas, *Tabula Rogeriana* or *The Recreation for Him Who Wishes to Travel Through the Countries*, in 1154. He incorporated the knowledge of Africa, the Indian Ocean and the Far East gathered by Arab merchants and explorers with the information inherited from the classical geographers to create the most accurate map of the world in pre-modern times. With funding from Roger II of Sicily (1097–1154), al-Idrisi drew on the knowledge collected at the University of Cordoba and paid draftsmen to make journeys and map their routes. The book describes the earth as a sphere with a circumference of 22,900 miles (36,900 km) but maps it in 70 rectangular sections. Notable features include the correct dual sources of the Nile, the coast of Ghana and mentions of Norway. Climate zones were a chief organizational principle. A second and shortened copy from 1192 called *Garden of Joys* is known by scholars as the *Little Idrisi*.

On the work of al-Idrisi, S. P. Scott commented:

The compilation of Edrisi marks an era in the history of science. Not only is its historical information most interesting and valuable, but its descriptions of many parts of the earth are still authoritative. For three centuries geographers copied his maps without alteration. The relative position of the lakes which form the Nile, as delineated in his work, does not differ greatly from that established by Baker and Stanley more than seven hundred years afterwards, and their number is the same. The mechanical genius of the author was not inferior to his erudition. The celestial and terrestrial planisphere of silver which he constructed for his royal patron was nearly six feet in diameter, and weighed four hundred and fifty pounds; upon the one side the zodiac and the constellations, upon the other—divided for convenience into segments—the bodies of land and water, with the respective situations of the various countries, were engraved.

— S. P. Scott , *History of the Moorish Empire*

1.1.1.7.4 **Piri Reis map**

The Ottoman cartographer Piri Reis published navigational maps in his *Kitab-ı Bahriye*. The work includes an atlas of charts for small segments of the Mediterranean, accompanied by the sailing instructions covering the sea. In the second version of the work, he included a map of the Americas. The Piri Reis map drawn by the Ottoman cartographer Piri Reis in 1513, is one of the oldest surviving maps show the Americas.

1.1.1.7.5 **Pacific islands**

The Polynesian peoples who explored and settled the Pacific islands in the first two millenniums AD used maps to navigate across large distances. A surviving map from the Marshall Islands uses sticks tied in a grid with palm strips representing wave and wind patterns, with shells attached to show the location of islands. Other maps were created as needed using temporary arrangements of stones or shells.

1.1.1.8 **European cartography**

1.1.1.8.1 **Medieval maps and the Mappa Mundi**

Medieval maps in Europe were mainly symbolic in form along the lines of the much earlier Babylonian World Map. Known as Mappa Mundi (cloth of the world) these maps were circular or symmetrical cosmological diagrams representing the Earth's single land mass as disk-shaped and surrounded by ocean.

1.1.1.8.2 **The Majorcan cartographic school and the Normal Portolan Chart**

The Majorcan cartographic school was a predominantly Jewish cooperation of cartographers, cosmographers and navigational instrument-makers in the late 13th to the 14th and 15th Century Majorca. With their multicultural heritage unstressed by fundamentalistic academic Christian traditions, the Majorcan cartographic school experimented and developed unique cartographic techniques. The Majorcan school was (co-) responsible for the invention (c. 1300) of the "Normal Portolan chart". It was a contemporary superiour, detailed nautical model chart, gridded by compass lines. The *Carta Pisana* portolan chart, made at the end of the 13th century (1275–1300), is the oldest surviving nautical chart (that is, not simply a map but a document showing accurate navigational directions).

1.1.1.8.3 **Roger Bacon and the Italian cartography school**

Roger Bacon's investigations of map projections and the appearance of portolano and then portolan charts for plying the European trade routes were rare innovations of the period. The Majorcan school is contrasted with the contemporary Italian map-making school.

1.1.1.8.4 **The Age of Exploration**

In the Renaissance, with the renewed interest in classical works, maps became more like surveys once again, while the discovery of the Americas by Europeans and the subsequent effort to control and divide those lands revived interest in scientific mapping methods. Peter Whitfield, the author of several books on the history of maps, credits European map-making as a factor in the global spread of western power: "Men in Seville, Amsterdam or London had access to knowledge of America, Brazil, or India, while the native peoples knew only their own immediate environment" (Whitfield). Jordan Branch and his advisor, Steven Weber, propose that the power of large kingdoms and nation states of later history are an inadvertent byproduct of 15th-century advances in map-making technologies.

- **15th century**: The monk Nicholas Germanus wrote a pioneering Cosmographia. He added the first new maps to Ptolemy's *Geographica*. Germanus invented the Donis map projection where the parallels of latitude are made equidistant, but meridians converge toward the poles.
- **c. 1485**: Portuguese cartographer Pedro Reinel made the oldest known signed Portuguese nautical chart.
- **1492**: German merchant Martin Behaim (1459–1507) made the oldest surviving terrestrial globe, but it lacked the Americas.
- **1492**: Cartographer Jorge de Aguiar made the oldest known signed and dated Portuguese nautical chart.

1.1.1.8.5 First maps of the Americas

The Spanish cartographer and explorer Juan de la Cosa sailed with Christopher Columbus. He created the first known cartographic representations showing both the Americas as well as Africa and Eurasia.

- **1502**: Unknown Portuguese cartographer made the Cantino planisphere, the first nautical chart to implicitly represent latitudes.
- **1504**: Portuguese cartographer Pedro Reinel made the oldest known nautical chart with a scale of latitudes.
- **1507**: Martin Waldseemüller's World Map was the first to use the term America for the Western continents (after explorer Amerigo Vespucci).
- **1519** : Portuguese cartographers Lopo Homem, Pedro Reinel and Jorge Reinel made the group of maps known today as the Miller Atlas or Lopo Homem – Reinéis Atlas.

1.1.1.8.6 **Diogo Ribeiro map (1527)**

Diogo Ribeiro, a Portuguese cartographer working for Spain, made what is considered the first scientific world map: the 1527 Padrón real The layout of the map (*Mapamundi*) is strongly influenced by the information obtained during the Magellan-Elcano trip around the world. Diogo's map delineates very precisely the coasts of Central and South America. The map shows, for the first time, the real extension of the Pacific Ocean. It also shows, for the first time, the North American coast as a continuous one (probably influenced by the Esteban Gómez's exploration in 1525). It also shows the demarcation of the

Treaty of Tordesillas.

1.1.1.8.7 **Gerardus Mercator (1569)**

Gerardus Mercator (1512–1594) was a Flemish cartographer who in his quest to make the world "look right" on the maps invented a new projection, called the Mercator projection. The projection was mathematically based and the Mercator maps gave much more accurate maps for world-wide navigation than any until that date. As in all cylindrical projections, parallels and meridians are straight and perpendicular to each other. In accomplishing this, the unavoidable east-west stretching of the map, is accompanied by a corresponding north-south stretching, so that at every point location, the east-west scale is the same as the north-south scale, making the projection conformal.

The development of the Mercator projection represented a major breakthrough in the nautical cartography of the 16th century. However, it was much ahead of its time, since the old navigational and surveying techniques were not compatible with its use in navigation. The Mercator projection would over time become the conventional view of the world that we are accustomed to today.

1.1.1.8.8 **Ortelius and the first atlas**

- **1570**: Antwerp cartographer Abraham Ortelius published the Theatrum Orbis Terrarum, the first modern atlas.
- **1608**: Captain John Smith published a map of Virginia's coastline.
- **1670s**: The astronomer Giovanni Domenico Cassini began work on the first modern topographic map in France. It was completed in 1789 or 1793 by his grandson Cassini de Thury.

1.1.1.8.9 **Enlightenment and scientific map-making**

- **1715**: Herman Moll published the Beaver Map, one of the most famous early maps of North America, which he copied from a 1698 work by Nicolas de Fer
- **1763–1767**: Captain James Cook mapped Newfoundland.

1.1.1.8.10 **Modern cartography**

The Vertical Perspective projection was first used by the German map publisher Matthias Seutter in 1740. He placed his observer at ~12,750 km distance. This is the type of projection used today by Google Earth.

The changes in the use of military maps was also part of the modern Military revolution, which changed the need for information as the scale of conflict increases as well. This created a need for maps to help with "...consistency, regularity and uniformity in military conflict."

The final form of the equidistant conic projection was constructed by the French astronomer Joseph-Nicolas Delisle in 1745.

The Swiss mathematician Johann Lambert invented several hemisperic map projections. In 1772 he created the Lambert conformal conic and Lambert azimuthal equal-area projections.

The Albers equal-area conic projection features no distortion along the standard parallels. It was invented by Heinrich Albers in 1805.

In the United States in the 17th and 18th centuries, explorers mapped trails and army engineers surveyed government lands. Two agencies were established to provide more detailed, large-scale mapping. They were the U.S. Geological Survey and the United States Coast and Geodetic Survey (now the National Geodetic Survey under the National Oceanic and Atmospheric Association).

The Greenwich prime meridian became the international standard reference for cartographers in 1884.

During the 20th century, maps became more abundant due to improvements in printing and photography that made production cheaper and easier. Airplanes made it possible to photograph large areas at a time.

Two-Point Equidistant projection was first drawn up by Hans Maurer in 1919. In this projection the distance from any point on the map to either of the two regulating points is accurate.

The loximuthal projection was constructed by Karl Siemon in 1935 and refined by Waldo Tobler in 1966.

Since the mid-1990s, the use of computers in map-making has helped to store, sort, and arrange data for mapping in order to create map projections.

1.1.2 **Technological changes**

In map-making, technology has continually changed in order to meet the demands of new generations of mapmakers and map users. The first maps were manually constructed with brushes and parchment and therefore varied in quality and were limited in distribution. The advent of the compass, printing press, telescope, sextant, quadrant and vernier allowed for the creation of far more accurate maps and the ability to make accurate reproductions. Professor Steven Weber of the University of California, Berkeley, has advanced the hypothesis that the concept of the "nation state" is an inadvertent byproduct of 15th-century advances in map-making technologies.

Advances in photochemical technology, such as the lithographic and photochemical processes, have allowed for the creation of maps that have fine details, do not distort in shape and resist moisture and wear. This also eliminated the need for engraving which further shortened the time it takes to make and reproduce maps.

In the mid-to-late 20th century, advances in electronic technology have led to further revolution in map-making. Specifically computer hardware devices such as computer screens, plotters, printers, scanners (remote and document) and analytic stereo plotters along with visualization, image processing, spatial analysis and database software, have democratized and greatly expanded the making of maps, particularly with their ability to produce maps that show slightly different features, without engraving a new printing plate.

1.1.3 **Map types**

1.1.3.1 **General vs thematic cartography**

In understanding basic maps, the field of map-making can be divided into two general categories: general cartography and thematic map-making. General map-making involves those maps that are constructed for a general audience and thus contain a variety of features. General maps exhibit many reference and

location systems and often are produced in a series. For example, the 1:24,000 scale topographic maps of the United States Geological Survey (USGS) are a standard as compared to the 1:50,000 scale Canadian maps. The government of the UK produces the classic 1:50,000 (replacing the older 1 inch to 1 mile) "Ordnance Survey" maps of the entire UK and with a range of correlated larger- and smaller-scale maps of great detail.

Thematic map-making involves maps of specific geographic themes, oriented toward specific audiences. A couple of examples might be a dot map showing corn production in Indiana or a shaded area map of Ohio counties, divided into numerical choropleth classes. As the volume of geographic data exploded over the last century, thematic map-making has become increasingly useful and necessary to interpret spatial, cultural and social data.

An orienteering map combines both general and thematic map-making, designed for a very specific user community. The most prominent theme element is shading, that indicates degrees of difficulty of travel due to vegetation. The vegetation itself is not identified, merely classified by the difficulty ("fight") that it presents.

1.1.3.2 **Topographic vs topological**

A topographic map is primarily concerned with the topographical description of a place, including (especially in the 20th and 21st centuries) the use of contour lines showing elevation. Terrain or relief can be shown in a variety of ways (see Cartographic relief depiction).

A topological map is a very general type of map, the kind you might sketch on a napkin. It often disregards scale and detail in the interest of clarity of communicating specific route or relational information. Beck's London Underground map is an iconic example. Though the most widely used maps of "The Tube," it preserves a little of reality: it varies scale constantly and abruptly, it straightens curved tracks, and it contorts directions. The only topography on it is the River Thames, letting the reader know whether a station is north or south of the river. That and the topology of station order and interchanges between train lines are all that is left of the geographic space. Yet those are all a typical passenger wishes to know, so the map fulfills its purpose.

1.1.3.3 **Map design**

1.1.4 **Map purpose and selection of information**

Arthur H. Robinson, an American cartographer influential in thematic map-making, stated that a map not properly designed "will be a cartographic failure." He also claimed, when considering all aspects of map-making, that "map design is perhaps the most complex." Robinson codified the mapmaker's understanding that a map must be designed foremost with consideration to the audience and its needs.

From the very beginning of map-making, maps "have been made for some particular purpose or set of purposes". The intent of the map should be illustrated in a manner in which the percipient acknowledges its purpose in a timely fashion. The term *percipient* refers to the person receiving information and was coined by Robinson. The principle of figure-ground refers to this notion of engaging the user by presenting a clear presentation, leaving no confusion concerning the purpose of the map. This will enhance the user's experience and keep his attention. If the user is unable to identify what is being demonstrated in a reasonable fashion, the map may be regarded as useless.

Making a meaningful map is the ultimate goal. Alan MacEachren explains that a well designed map "is convincing because it implies authenticity" (1994, pp. 9). An interesting map will no doubt engage a reader. Information richness or a map that is multivariate shows relationships within the map. Showing several variables allow comparison, which adds to the meaningfulness of the map. This also generates hypothesis and stimulates ideas and perhaps further research. In order to convey the message of the map, the creator must design it in a manner which will aid the reader in the overall understanding of its purpose. The title of a map may provide the "needed link" necessary for communicating that message, but the overall design of the map fosters the manner in which the reader interprets it (Monmonier, 1993, pp. 93).

In the 21st century it is possible to find a map of virtually anything from the inner workings of the human body to the virtual worlds of cyberspace. Therefore there are now a huge variety of different styles and types of map - for example, one area which has evolved a specific and recognizable variation are those used by public transport organizations to guide passengers, namely urban rail and metro maps, many of which are loosely based on 45 degree angle as originally perfected by Harry Beck and George Dow.

1.1.5 **Naming conventions**

Most maps use text to label places and for such things as the map title, legend and other information. Although maps are often made in one specific language, place names often differ between languages. So a map made in English may use the name *Germany* for that country, while a German map would use *Deutschland* and a French map *Allemagne*. A non-native term for a place is referred to as an exonym.

In some cases the correct name is not clear. For example, the nation of Burma officially changed its name to Myanmar, but many nations do not recognize the ruling junta and continue to use *Burma*. Sometimes an official name change is resisted in other languages and the older name may remain in common use. Examples include the use of *Saigon* for Ho Chi Minh City, *Bangkok* for Krung Thep and *Ivory Coast* for Côte d'Ivoire.

Difficulties arise when transliteration or transcription between writing systems is required. Some well-known places have well-established names in other languages and writing systems, such as *Russia* or *Rußland* for Росси́я, but in other cases a system of transliteration or transcription is required. Even in the former case, the exclusive use of an exonym may be unhelpful for the map user. It will not be much use for an English user of a map of Italy to show Livorno *only* as "Leghorn" when road signs and railway timetables show it as "Livorno". In transliteration, the characters in one script are represented by characters in another. For example, the Cyrillic letter *P* is usually written as *R* in the Latin script, although in many cases it is not as simple as a one-for-one equivalence. Systems exist for transliteration of Arabic, but the results may vary. For example, the Yemeni city of Mocha is written variously in English as Mocha, Al Mukha, al-Mukhā, Mocca and Moka. Transliteration systems are based on relating written symbols to one another, while transcription is the attempt to spell in one language the phonetic sounds of another. Chinese writing is now usually converted to the Latin alphabet through the Pinyin phonetic transcription systems. Other systems were used in the past, such as Wade-Giles, resulting in the city being spelled *Beijing* on newer English maps and *Peking* on older ones.

Further difficulties arise when countries, especially former colonies, do not have a strong national geographic naming standard. In such cases, cartographers may have to choose between various phonetic spellings of local names versus older imposed, sometimes resented, colonial names. Some countries have multiple official languages, resulting in multiple official place names. For example, the capital of Belgium is both *Brussel* and *Bruxelles*. In Canada, English and French are official languages and places

have names in both languages. British Columbia is also officially named *la Colombie-Britannique*. English maps rarely show the French names outside of Quebec, which itself is spelled *Québec* in French.

The study of place names is called toponymy, while that of the origin and historical usage of place names as words is etymology.

In order to improve legibility or to aid the illiterate, some maps have been produced using pictograms to represent places. The iconic example of this practice is Lance Wyman's early plans for the Mexico City Metro, on which stations were shown simply as stylized logos. Wyman also prototyped such a map for the Washington Metro, though ultimately the idea was rejected. Other cities experimenting with such maps are Fukuoka, Guadalajara and Monterrey.

1.1.6 **Map symbology**

The quality of a map's design affects its reader's ability to extract information and to learn from the map. Cartographic symbology has been developed in an effort to portray the world accurately and effectively convey information to the map reader. A legend explains the pictorial language of the map, known as its symbology. The title indicates the region the map portrays; the map image portrays the region and so on. Although every map element serves some purpose, convention only dictates inclusion of some elements, while others are considered optional. A menu of map elements includes the neat line (border), compass rose or north arrow, overview map, bar scale, map projection and information about the map sources, accuracy and publication.

When examining a landscape, scale can be intuited from trees, houses and cars. Not so with a map. Even such a simple thing as a north arrow is crucial. It may seem obvious that the top of a map should point north, but this might not be the case.

Map coloring is also very important. How the cartographer displays the data in different hues can greatly affect the understanding or feel of the map. Different intensities of hue portray different objectives the cartographer is attempting to get across to the audience. Today, personal computers can display up to 16 million distinct colors at a time. This fact allows for a multitude of color options for even for the most demanding maps. Moreover, computers can easily hatch patterns in colors to give even more options. This is very beneficial, when symbolizing data in categories such as quintile and equal interval classifications.

Quantitative symbols give a visual measure of the relative size/importance/number that a symbol represents and to symbolize this data on a map, there are two major classes of symbols used for portraying quantitative properties. Proportional symbols change their visual weight according to a quantitative property. These are appropriate for extensive statistics. Choropleth maps portray data collection areas, such as counties or census tracts, with color. Using this color way, the darkness and intensity (or value) of the color are evaluated by the eye as a measure of intensity or concentration.

1.1.7 **Cartographic relief depiction**

Terrain or relief is an essential aspect of physical geography, and as such its portrayal presents a central problem in map-making, and more recently GIS and 3D Visualization.

The most obvious way to depict relief is to imitate it at scale, as in molded or sculpted solid terrain models and molded-plastic raised-relief maps. Because of the disparity between the horizontal and vertical scales of maps, raised relief is typically exaggerated.

1.1.7.1 **Hill profiles**

The most ancient form of relief depiction in map-making, **hill profiles** are simply illustrations of mountains and hills in profile, placed as appropriate on generally small-scale (broad area of coverage) maps. They are seldom used today except as part of an "antique" styling.

1.1.7.2 **Hachures**

Hachures are also an older mode of representing relief. They are a form of shading, although different from the one used in shaded maps. They show the orientation of slope, and by their thickness and overall density, they provide a general sense of the steepness. Being non-numeric, they are less useful to a scientific survey than contours, but can successfully communicate quite specific shapes of terrain.

Hachure representation of relief was standardized by the Austrian topographer Johann Georg Lehmann in 1799.

1.1.8 **Hypsometric tints**

Hypsometric tints are related to contour lines. They can be used to depict ranges of elevation as bands of color, usually in a graduated scheme, or as a color ramp applied to contour lines themselves. A typical scheme progresses from dark greens for lower elevations up through yellows/browns, and on to grays and white at the highest elevations. Hypsometric tinting of maps and globes is often accompanied by a similar method of bathymetric tinting to convey depth of the oceans; lighter shades of blue represent shallower water such as the continental shelf and darker shade deeper regions.

1.1.9 **Shaded relief**

Shaded relief, or hill-shading, simulates the cast shadow thrown upon a raised relief map, or more abstractly upon the planetary surface represented. The shadows normally follow the English convention of top-left lighting in which the light source is placed near the upper-left corner of the map. If the map is oriented with north at the top, the result is that the light appears to come from the north-west. Many people have pointed out that this is unrealistic for maps of the northern hemisphere, because the sun does not shine from that direction, and they have proposed using southern lighting. However, the normal convention is followed to avoid multistable perception illusions (i.e. crater/hill confusion).

Traditionally drawn with charcoal, airbrush and other artist's media, shaded relief is today almost exclusively computer-generated using digital elevation models, with a resulting different look and feel. Much work has been done in digitally recreating the work of Swiss master Eduard Imhof, widely regarded as the master of manual hill-shading technique and theory. Imhof's contributions included a multi-color approach to shading, with purples in valleys and yellows on peaks.

The use of illumination and shadow to produce an appearance of three-dimensional space on a flat-surfaced map closely parallels the painting technique known as chiaroscuro.

Shaded relief today can be created digitally, using a digital elevation model (DEM) as its basis. The DEM may be converted to shaded relief using software such as Photoshop or ArcMap's Spatial Analyst extension.

1.1.10 **Physiographic illustration**

Pioneered by Hungarian-American cartographer Erwin Raisz, this technique uses generalized texture to imitate landform shapes over a large area. A combination of hill profile and shaded relief, this style of relief representation is simultaneously idiosyncratic to its creator and very useful in illustrating geomorphological patterns.

More recently, Tom Patterson created a computer-generated map of the United States using Erwin Raisz's work as a starting point, the *Physical Map of the Coterminous United States*

1.1.11 **Forums and associations**

Portrayal of relief is especially important in mountainous regions. The Commission on Mountain Cartography of the International Cartographic Association is the best-known forum for discussion of theory and techniques for mapping these regions.

1.2 Two Dimensional Diagrams

A **diagram** is a two-dimensional geometric symbolic representation of information according to some visualization technique. Sometimes, the technique uses a three-dimensional visualization which is then projected onto the two-dimensional surface. The word *graph* is sometimes used as a synonym for diagram.

1.2.1 **Overview**

The term diagram in common sense can have a general or specific meaning:

- *Visual information device* : Like the term "illustration" the diagram is used as a collective term standing for the whole class of technical genres, including graphs, technical drawings and tables.
- *Specific kind of visual display* : This is only the genre, that show qualitative data with shapes that are connected by lines, arrows, or other visual links.

In science the term is used in both ways. For example Anderson (1997) stated more generally: "diagrams are pictorial, yet abstract, representations of information, and maps, line graphs, bar charts, engineering blueprints, and architects' sketches are all examples of diagrams, whereas photographs and video are not". On the other hand Lowe (1993) defined diagrams as specifically "abstract graphic portrayals of the subject matter they represent".

In the specific sense diagrams and charts contrast computer graphics, technical illustrations, infographics, maps, and technical drawings, by showing "abstract rather than literal representations of information". The essences of a diagram can be seen as:

- A *form* of visual formatting devices
- A *display* that does not show quantitative data, but rather relationships and abstract information
- With *building blocks* such as geometrical shapes connected by lines, arrows, or other visual links.

Or in Hall's (1996) words "diagrams are simplified figures, caricatures in a way, intended to convey essential meaning". These simplified figures are often based on a set of rules. The basic shape according to White (1984) can be characterized in terms of "elegance, clarity, ease, pattern, simplicity, and validity".

The elegance for a start is determined by whether or not the diagram is "the simplest and most fitting solution to a problem".

1.2.2 **Main diagram types**

There are at least the following types of diagrams:

- Graph-based diagrams: these take a collection of items and relationships between them, and express them by giving each item a 2D position, while the relationships are expressed as connections between the items or overlaps between the items.
- Chart-like diagram techniques, which display a relationship between two variables that take either discrete or a continuous range of values.
- Schematics and other types of diagrams.

1.3 Divided Rectangle

In Euclidean plane geometry, a **rectangle** is any quadrilateral with four right angles. It can also be defined as an equiangular quadrilateral, since equiangular means that all of its angles are equal (360°/4 = 90°). It can also be defined as a parallelogram containing a right angle. A rectangle with four sides of equal length is a square. The term **oblong** is occasionally used to refer to a non-square rectangle. A rectangle with vertices *ABCD* would be denoted as $\square ABCD$.

The word rectangle comes from the Latin *rectangulus*, which is a combination of *rectus* (right) and *angulus* (angle).

A so-called **crossed rectangle** is a crossed (self-intersecting) quadrilateral which consists of two opposite sides of a rectangle along with the two diagonals. It is a special case of an antiparallelogram, and its angles are not right angles. Other geometries, such as spherical, elliptic, and hyperbolic, have so-called rectangles with opposite sides equal in length and equal angles that are not right angles.

Rectangles are involved in many tiling problems, such as tiling the plane by rectangles or tiling a rectangle by polygons.

1.3 .1 **Characterizations**

A convex quadrilateral is a rectangle iff (if and only if) it is any one of the following:

- A parallelogram with at least one right angle
- A quadrilateral with four right angles
- Equiangular
- A parallelogram with diagonals of equal length
- A parallelogram *ABCD* where triangles *ABD* and *DCA* are congruent
- A convex quadrilateral with successive sides *a*, *b*, *c*, *d* whose area is $\frac{1}{2}\sqrt{(a^2+c^2)(b^2+d^2)}$.
- A convex quadrilateral with successive sides *a*, *b*, *c*, *d* whose area is $\frac{1}{4}(a+c)(b+d)$.

1.3.2 **Classification**

1.3.2.1 **Traditional hierarchy**

A rectangle is a special case of a parallelogram in which each pair of adjacent sides is perpendicular.

A parallelogram is a special case of a trapezium (known as a trapezoid in North America) in which *both* pairs of opposite sides are parallel and equal in length.

A trapezium is a convex quadrilateral which has at least one pair of parallel opposite sides.

A convex quadrilateral is

- **Star-shaped**: The whole interior is visible from a single point, without crossing any edge.
- **Simple**: The boundary does not cross itself.

1.3.2.2 **Alternative hierarchy**

De Villiers defines a rectangle more generally as any quadrilateral with axes of symmetry through each pair of opposite sides. This definition includes both right-angled rectangles and crossed rectangles. Each has an axis of symmetry parallel to and equidistant from a pair of opposite sides, and another which is the perpendicular bisector of those sides, but, in the case of the crossed rectangle, the first axis is not an axis of symmetry for either side that it bisects.

Quadrilaterals with two axes of symmetry, each through a pair of opposite sides, belong to the larger class of quadrilaterals with at least one axis of symmetry through a pair of opposite sides. These quadrilaterals comprise isosceles trapezia and crossed isosceles trapezia (crossed quadrilaterals with the same vertex arrangement as isosceles trapezia).

1.3.3 **Properties**

1.3.3.1 **Symmetry**

A rectangle is cyclic: all corners lie on a single circle.

It is equiangular: all its corner angles are equal (each of 90 degrees).

It is isogonal or vertex-transitive: all corners lie within the same symmetry orbit.

It has two lines of reflectional symmetry and rotational symmetry of order 2 (through 180°).

1.3.3.2 **Rectangle-rhombus duality**

The dual polygon of a rectangle is a rhombus, as shown in the table below.

Rectangle	**Rhombus**
All *angles* are equal.	All *sides* are equal.
Alternate *sides* are equal.	Alternate *angles* are equal.

Its center is equidistant from its *vertices*, hence it has a *circumcircle*.	Its center is equidistant from its *sides*, hence it has an *in-circle*.
Its axes of symmetry bisect opposite *sides*.	Its axes of symmetry bisect opposite *angles*.
Diagonals are equal in *length*.	Diagonals intersect at equal *angles*.

Miscellaneous

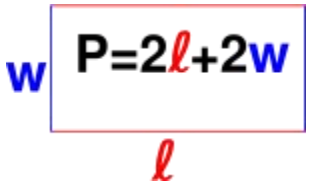
The two diagonals are equal in length and bisect each other. Every quadrilateral with both these properties is a rectangle.

A rectangle is rectilinear: its sides meet at right angles.

A rectangle in the plane can be defined by five independent degrees of freedom consisting, for example, of three for the position (comprising two of translation and one of rotation), one for shape (aspect ratio), and one for overall size (area).

Two rectangles, neither of which will fit inside the other, are said to be incomparable.

Formulae



The formula for the perimeter of a rectangle.

If a rectangle has length ℓ and width *w*

- It has area $K = \ell w$,
- It has perimeter $P = 2\ell + 2w = 2(\ell + w)$,
- Each diagonal has length $d = \sqrt{\ell^2 + w^2}$,
- And when $\ell = w$, the rectangle is a square.

1.3.4 **Theorems**

The isoperimetric theorem for rectangles states that among all rectangles of a given perimeter, the square has the largest area.

The midpoints of the sides of any quadrilateral with perpendicular diagonals form a rectangle.

A parallelogram with equal diagonals is a rectangle.

The Japanese theorem for cyclic quadrilaterals states that the incentres of the four triangles determined by the vertices of a cyclic quadrilateral taken three at a time form a rectangle.

The British flag theorem states that with vertices denoted *A*, *B*, *C*, and *D*, for any point *P* on the same plane of a rectangle:

$$(AP)^2 + (CP)^2 = (BP)^2 + (DP)^2.$$

1.3.5 **Crossed rectangles**

A crossed (self-intersecting) quadrilateral consists of two opposite sides of a non-self-intersecting quadrilateral along with the two diagonals. Similarly, a crossed rectangle is a crossed quadrilateral which consists of two opposite sides of a rectangle along with the two diagonals. It has the same vertex arrangement as the rectangle. It appears as two identical triangles with a common vertex, but the geometric intersection is not considered a vertex.

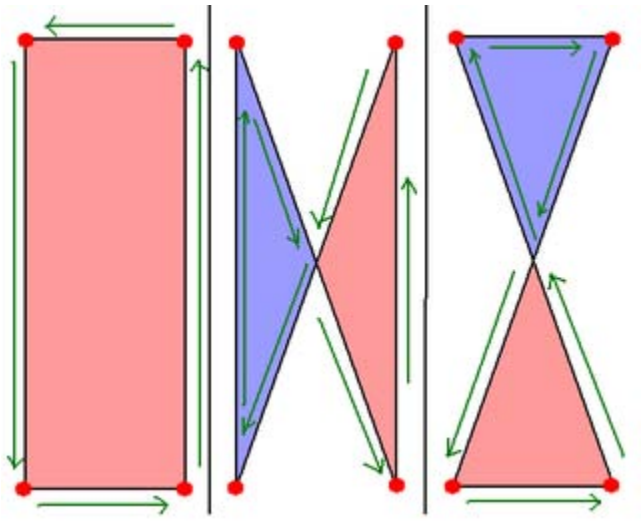
A crossed quadrilateral is sometimes likened to a bow tie or butterfly. A three-dimensional rectangular wire frame that is twisted can take the shape of a bow tie. A crossed rectangle is sometimes called an "angular eight".

The interior of a crossed rectangle can have a polygon density of ±1 in each triangle, dependent upon the winding orientation as clockwise or counterclockwise.

A crossed rectangle is not equiangular. The sum of its interior angles (two acute and two reflex), as with any crossed quadrilateral, is 720°.

A rectangle and a crossed rectangle are quadrilaterals with the following properties in common:

- Opposite sides are equal in length.
- The two diagonals are equal in length.
- It has two lines of reflectional symmetry and rotational symmetry of order 2 (through 180°).



1.3.6 **Other rectangles**

In solid geometry, a figure is non-planar if it is not contained in a (flat) plane. A **skew rectangle** is a non-planar quadrilateral with opposite sides equal in length and four equal acute angles. A **saddle rectangle** is

a *skew rectangle* with vertices that alternate an equal distance above and below a plane passing through its center, named for its minimal surface interior seen with saddle point at its center. The convex hull of this skew rectangle is a special tetrahedron called a rhombic disphenoid. (The term "skew rectangle" is also used in 2D graphics to refer to a distortion of a rectangle using a "skew" tool. The result can be a parallelogram or a trapezoid/trapezium.)

In spherical geometry, a **spherical rectangle** is a figure whose four edges are great circle arcs which meet at equal angles greater than 90°. Opposite arcs are equal in length. The surface of a sphere in Euclidean solid geometry is a non-Euclidean surface in the sense of elliptic geometry. Spherical geometry is the simplest form of elliptic geometry.

In elliptic geometry, an **elliptic rectangle** is a figure in the elliptic plane whose four edges are elliptic arcs which meet at equal angles greater than 90°. Opposite arcs are equal in length.

In hyperbolic geometry, a **hyperbolic rectangle** is a figure in the hyperbolic plane whose four edges are hyperbolic arcs which meet at equal angles less than 90°. Opposite arcs are equal in length.

1.3.7 **Squared, perfect, and other tiled rectangles**

A rectangle tiled by squares, rectangles, or triangles is said to be a "squared", "rectangled", or "triangulated" (or "triangled") rectangle respectively. The tiled rectangle is *perfect* if the tiles are similar and finite in number and no two tiles are the same size. If two such tiles are the same size, the tiling is *imperfect*. In a perfect (or imperfect) triangled rectangle the triangles must be right triangles.

A rectangle has commensurable sides if and only if it is tileable by a finite number of unequal squares. The same is true if the tiles are unequal isosceles right triangles.

The tilings of rectangles by other tiles which have attracted the most attention are those by congruent non-rectangular polyominoes, allowing all rotations and reflections. There are also tilings by congruent polyaboloes.

1.4 **Flowchart**

- A flowchart is a type of diagram that represents an algorithm or process, showing the steps as boxes of various kinds, and their order by connecting them with arrows. This diagrammatic representation illustrates a solution to a given problem. Process operations are represented in these boxes, and arrows; rather, they are implied by the sequencing of operations. Flowcharts are used in analyzing, designing, documenting or managing a process or program in various fields.

1.4.1 Overview

Flowcharts are used in designing and documenting complex processes or programs. Like other types of diagrams, they help visualize what is going on and thereby help the viewer to understand a process, and perhaps also find flaws, bottlenecks, and other less-obvious features within it. There are many different types of flowcharts, and each type has its own repertoire of boxes and notational conventions. The two most common types of boxes in a flowchart are:

- A processing step, usually called *activity*, and denoted as a rectangular box
- A decision, usually denoted as a diamond.

A flow chart is described as "cross-functional" when the page is divided into different swimlanes describing the control of different organizational units. A symbol appearing in a particular "lane" is within the control of that organizational unit. This technique allows the author to locate the responsibility for performing an action or making a decision correctly, showing the responsibility of each organizational unit for different parts of a single process.

Flowcharts depict certain aspects of processes and they are usually complemented by other types of diagram. For instance, Kaoru Ishikawa defined the flowchart as one of the seven basic tools of quality control, next to the histogram, Pareto chart, check sheet, control chart, cause-and-effect diagram, and the scatter diagram. Similarly, in UML, a standard concept-modeling notation used in software development, the activity diagram, which is a type of flowchart, is just one of many different diagram types.

Nassi-Shneiderman diagrams are an alternative notation for process flow.

Common alternate names include: flowchart, process flowchart, functional flowchart, process map, process chart, functional process chart, business process model, process model, process flow diagram, work flow diagram, business flow diagram. The terms "flow chart" and "flow chart" are used interchangeably.

1.4.2 **History**

The first structured method for documenting process flow, the "flow process chart", was introduced by Frank Gilbreth to members of the American Society of Mechanical Engineers (ASME) in 1921 in the presentation “Process Charts—First Steps in Finding the One Best Way”. Gilbreth's tools quickly found their way into industrial engineering curricula. In the early 1930s, an industrial engineer, Allan H. Mogensen began training business people in the use of some of the tools of industrial engineering at his Work Simplification Conferences in Lake Placid, New York.

A 1944 graduate of Mogensen's class, Art Spinanger, took the tools back to Procter and Gamble where he developed their Deliberate Methods Change Program. Another 1944 graduate, Ben S. Graham, Director of Formcraft Engineering at Standard Register Industrial, adapted the flow process chart to information processing with his development of the multi-flow process chart to display multiple documents and their relationships. In 1947, ASME adopted a symbol set derived from Gilbreth's original work as the ASME Standard for Process Charts.

Douglas Hartree explains that Herman Goldstine and John von Neumann developed a flow chart (originally, diagram) to plan computer programs. His contemporary account is endorsed by IBM engineers and by Goldstine's personal recollections. The original programming flowcharts of Goldstine and von Neumann can be seen in their unpublished report, "Planning and coding of problems for an electronic computing instrument, Part II, Volume 1" (1947), which is reproduced in von Neumann's collected works.

Flowcharts used to be a popular means for describing computer algorithms and are still used for this purpose. Modern techniques such as UML activity diagrams can be considered to be extensions of the flowchart. In the 1970s the popularity of flowcharts as an own method decreased when interactive computer terminals and third-generation programming languages became the common tools of the trade, since algorithms can be expressed much more concisely as source code in such a language, and also because designing algorithms using flowcharts was more likely to result in spaghetti code because of the need for gotos to describe arbitrary jumps in control flow. Often the pseudo - code is used, which uses the common idioms of such languages without strictly adhering to the details of a particular one.

1.4.3 **Flowchart building blocks**

1.4.3.1 **Examples**

A flow chart for computing the factorial of N — written N! and equal to $1 \times 2 \times 3 \times ... \times N$.

1.4.4 **Symbols**

Flow Chart Symbols List A typical flowchart from older basic computer science textbooks may have the following kinds of symbols:

Start and end symbols
: Represented as circles, ovals or rounded (fillet) rectangles, usually containing the word "Start" or "End", or another phrase signaling the start or end of a process, such as "submit inquiry" or "receive product".

Arrows
: Showing "flow of control". An arrow coming from one symbol and ending at another symbol represents that control passes to the symbol the arrow points to. The line for the arrow can be solid or dashed. The meaning of the arrow with dashed line may differ from one flowchart to another and can be defined in the legend.

Generic processing steps
: Represented as rectangles. Examples: "Add 1 to X"; "replace identified the part"; "save changes" or similar.

Subroutines
: Represented as rectangles with double-struck vertical edges; these are used to show complex processing steps which may be detailed in a separate flow chart. Example: PROCESS-FILES. One subroutine may have multiple distinct entry points or exit flows (see coroutine); if so, these are shown as labeled 'wells' in the rectangle, and control arrows connect to these 'wells'.

Input/Output
: Represented as a parallelogram. Examples: Get X from the user; display X.

Prepare conditional
: Represented as a hexagon. Shows operations which have no effect other than preparing a value for a subsequent conditional or decision step (see below).

Conditional or decision
: Represented as a diamond (rhombus) showing where a decision is necessary, commonly a Yes/No question or True/False test. The conditional symbol is peculiar in that it has two arrows coming out of it, usually from the bottom point and the right point, one corresponding to Yes or True, and one corresponding to No or False. (The arrows should always be labeled.) More than two arrows can be used, but this is normally a clear indicator that a complex decision is being taken, in which case it may need to be broken-down further or replaced with the "pre-defined process" symbol.

Junction symbol
: Generally represented with a black blob, showing where multiple control flows converge in a single exit flow. A junction symbol will have more than one arrow coming into it, but only one going out.
: In simple cases, one may simply have an arrow point to another arrow instead. These are useful to represent an iterative process (what in Computer Science is called a loop). A loop may, for example, consist of a connector where control first enters, processing steps, a conditional with one arrow exiting the loop, and one going back to the connector.

For additional clarity, wherever two lines accidentally cross in the drawing, one of them may be drawn with a small semicircle over the other, showing that no junction is intended.

Labeled connectors

Represented by an identifying label inside a circle. Labeled connectors are used in complex or multi-sheet diagrams to substitute for arrows. For each label, the "outflow" connector must always be unique, but there may be any number of "inflow" connectors. In this case, a junction in control flow is implied.

Concurrency symbol

Represented by a double transverse line with any number of entry and exit arrows. These symbols are used whenever two or more control flows must operate simultaneously. The exit flows are activated concurrently when all of the entry flows have reached the concurrency symbol. A concurrency symbol with a single entry flow is a *fork*; one with a single exit flow is a *join*.

All processes should flow from top to bottom and left to right.

1.4.5 **Data-flow extensions**

A number of symbols have been standardized for data flow diagrams represent data flow, rather than control flow. These symbols may also be used in control flowcharts (e.g. to substitute for the parallelogram symbol).

- A *Document* represented as a rectangle with a wavy base;
- A *Manual input* represented by quadrilateral, with the top irregularly sloping up from left to right. An example would be to signify data-entry from a form;
- A *Manual operation* represented by a trapezoid with the longer parallel side at the top, to represent an operation or adjustment process that can only be made manually.
- A *Data File* represented by a cylinder.

1.4.6 **Types of flowchart**

Sterneckert (2003) suggested that flowcharts can be modeled from the perspective of different user groups (such as managers, system analysts and clerks) and that there are four general types:

- *Document flowcharts*, showing controls over a document-flow through a system
- *Data flowcharts*, showing controls over a data-flow in a system
- *System flowcharts* showing controls at a physical or resource level
- *Program flowchart*, showing the controls in a program within a system

Notice that every type of flowchart focuses on some kind of control, rather than on the particular flow itself.

However there are several of these classifications. For example Andrew Veronis (1978) named three basic types of flowcharts: the *system flow chart*, the *general flowchart*, and the *detailed flowchart*. That same year Marilyn Bohl (1978) stated "in practice, two kinds of flowcharts are used in solution planning: *system flow charts* and *program flowcharts*...". More recently Mark A. Fryman (2001) stated that there are more differences: "Decision flow charts, logic flow charts, systems flowcharts, product flow charts, and process flowcharts are just a few of the different types of flowcharts that are used in business and government".

In addition, many diagram techniques exist that are similar to flowcharts but carry a different name, such as UML activity diagrams.

1.5 Square Diagram

The ***N*² chart**, also referred to as ***N*² diagram**, ***N*-squared diagram** or ***N*-squared chart**, is a diagram in the shape of a matrix, representing functional or physical interfaces between system elements. It is used to systematically identify, define, tabulate, design, and analyze functional and physical interfaces. It applies to system interfaces and hardware and/or software interfaces.

The *N*-squared chart was invented by the systems engineer Robert J. Lano, while working at TRW in the 1970s and first published in a 1977 TRW internal report.

1.5.1 **Overview**

The N^2 diagram has been used extensively to develop data interfaces, primarily in the software areas. However, it can also be used to develop hardware interfaces. The basic N^2 chart is shown in Figure 2. The system functions are placed on the diagonal; the remainder of the squares in the $N \times N$ matrix represents the interface inputs and outputs.

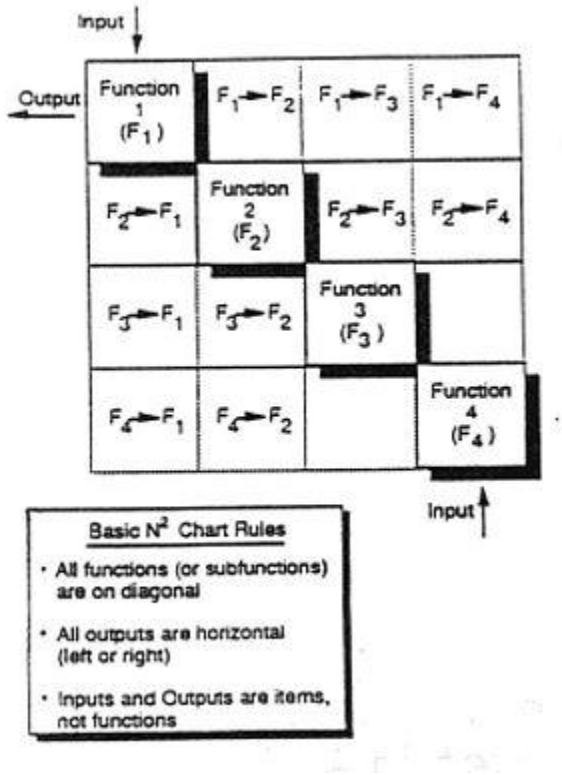


Figure 2. N2 chart definition.

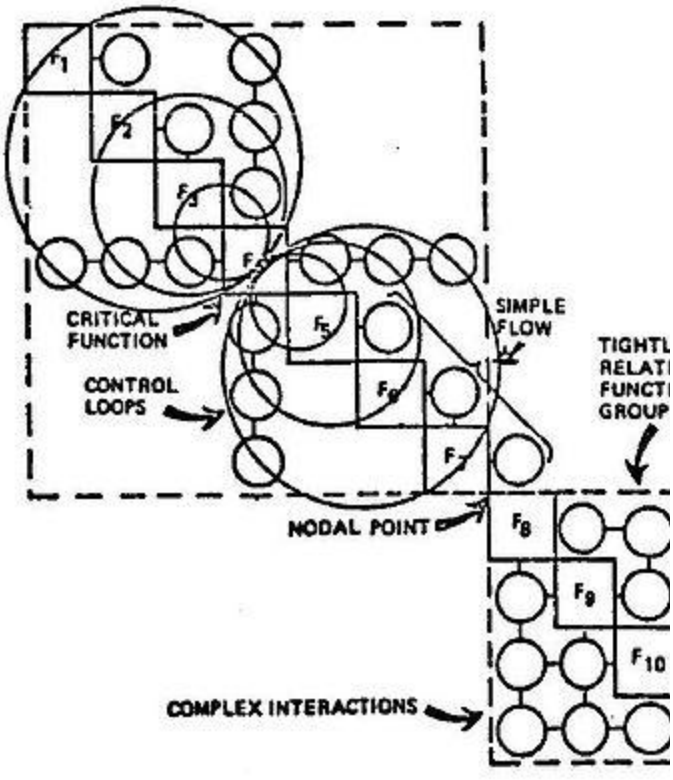


Figure 3. N2 Chart Key Features.

Where a blank appears, there is no interface between the respective functions. Data flows in a clockwise direction between functions (e.g., the symbol F1 F2 indicates data flowing from function F1, to function F2). The data being transmitted can be defined in the appropriate squares. Alternatively, the use of circles

and numbers permits a separate listing of the data interfaces. The clockwise flow of data between functions that have a feedback loop can be illustrated by a larger circle called a control loop. The identification of a critical function is also shown in Figure 3, where function F4 has a number of inputs and outputs to all other functions in the upper module. A simple flow of interface data exists between the upper and lower modules at functions F7 and F8. The lower module has complex interaction among its functions. The N2 chart can be taken down into successively lower levels to the hardware and software component functional levels. In addition to defining the data that must be supplied across the interface, the N2 chart can pinpoint areas where conflicts could arise.

1.5.2 N^2 charts building blocks

1.5.2.1 Number of entities

The "*N*" in an N^2 diagram is the number of entities for which relationships are shown. This $N \times N$ matrix requires the user to generate complete definitions of all interfaces in a rigid bidirectional, fixed framework. The user places the functional or physical entities on the diagonal axis and the interface inputs and outputs in the remainder of the diagram squares. A blank square indicates that there is no interface between the respective entities. Data flows clockwise between entities (i.e., the symbol F1 ? F2 in Figure 1 indicates data flowing from function F1 to function F2; the symbol F2 = F1 indicates the feedback). That which passes across the interface is defined in the appropriate squares.

The diagram is complete when the user has compared each entity to all other entities. The N2 diagram should be used in each successively lower level of entity decomposition. Figure 1 illustrates the directional flow of interfaces between entities within an N^2 diagram. (In this case, the entities are functions.)

1.5.2.2 **Functions on the diagonal**

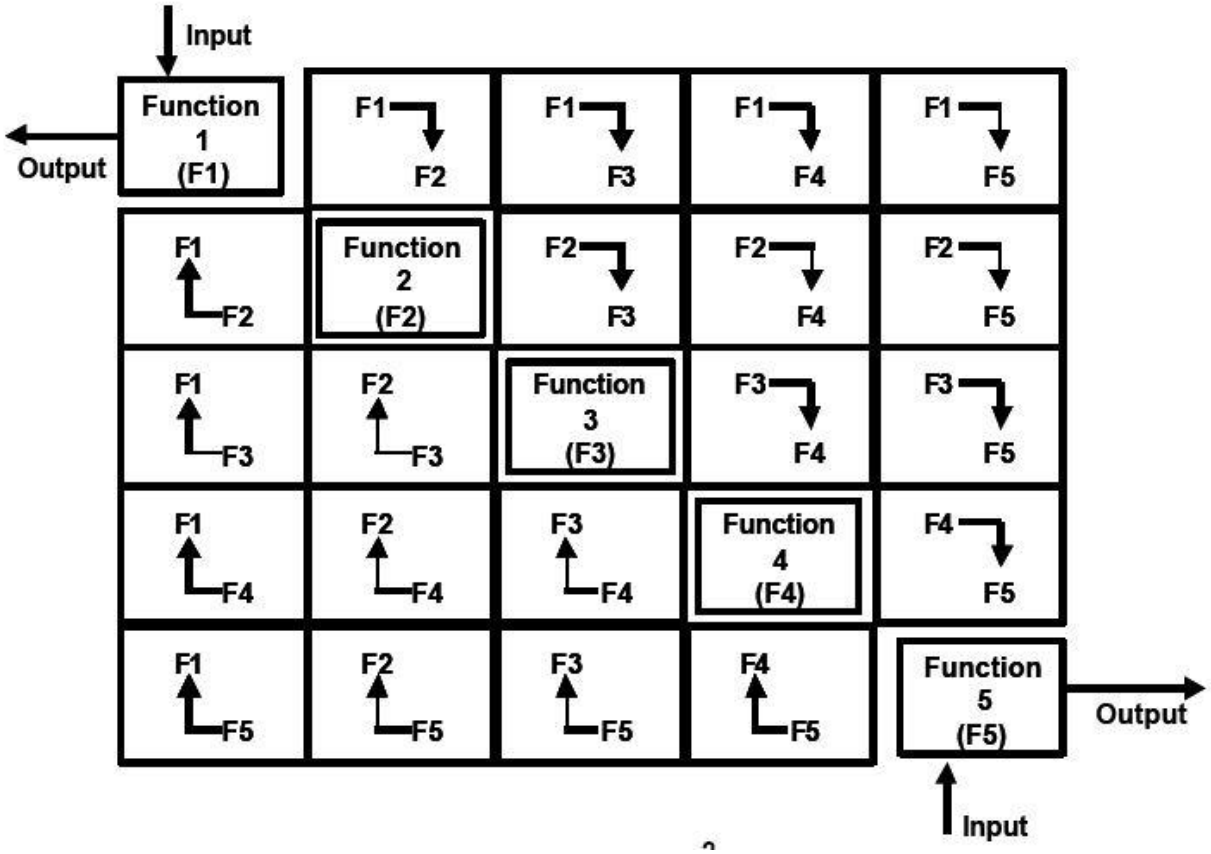


Figure 4. N^2 diagram.

In the example on the right, *N* equals 5. The five functions are on the diagonal. The arrows show the flow of data between functions. So if function 1 sends data to function 2, the data elements would be placed in the box to the right of function 1. If function 1 does not send data to any of the other functions, the rest of the boxes to right of function 1 would be empty. If function 2 sends data to function 3 and function 5, then the data elements would be placed in the first and third boxes to the right of function 2. If any function sends data back to a previous function, then the associated box to the left of the function would have the data elements placed in it. The squares on either side of the diagonal (not just adjacent squares) are filled with appropriate data to depict the flow between the functions. If there is no interface between two functions, the square that represents the interface between the two functions is left blank. Physical interfaces would be handled in the same manner, with the physical entities on the diagonal rather than the functional entities.

1.5.2.3 **Contextual and administrative data**

Each N^2 diagram shall contain at a minimum the following contextual and administrative data:

- Date the diagram was created
- Name of the engineer, organization, or working group that created the diagram
- A Unique decimal delimited number of the functional or physical entity being diagrammed
- Unique name of the functional or physical entity being diagrammed

N2 diagrams are a valuable tool for not only identifying functional or physical interfaces, but also for pinpointing areas in which conflicts may arise with interfaces so that system integration proceeds smoothly and efficiently.

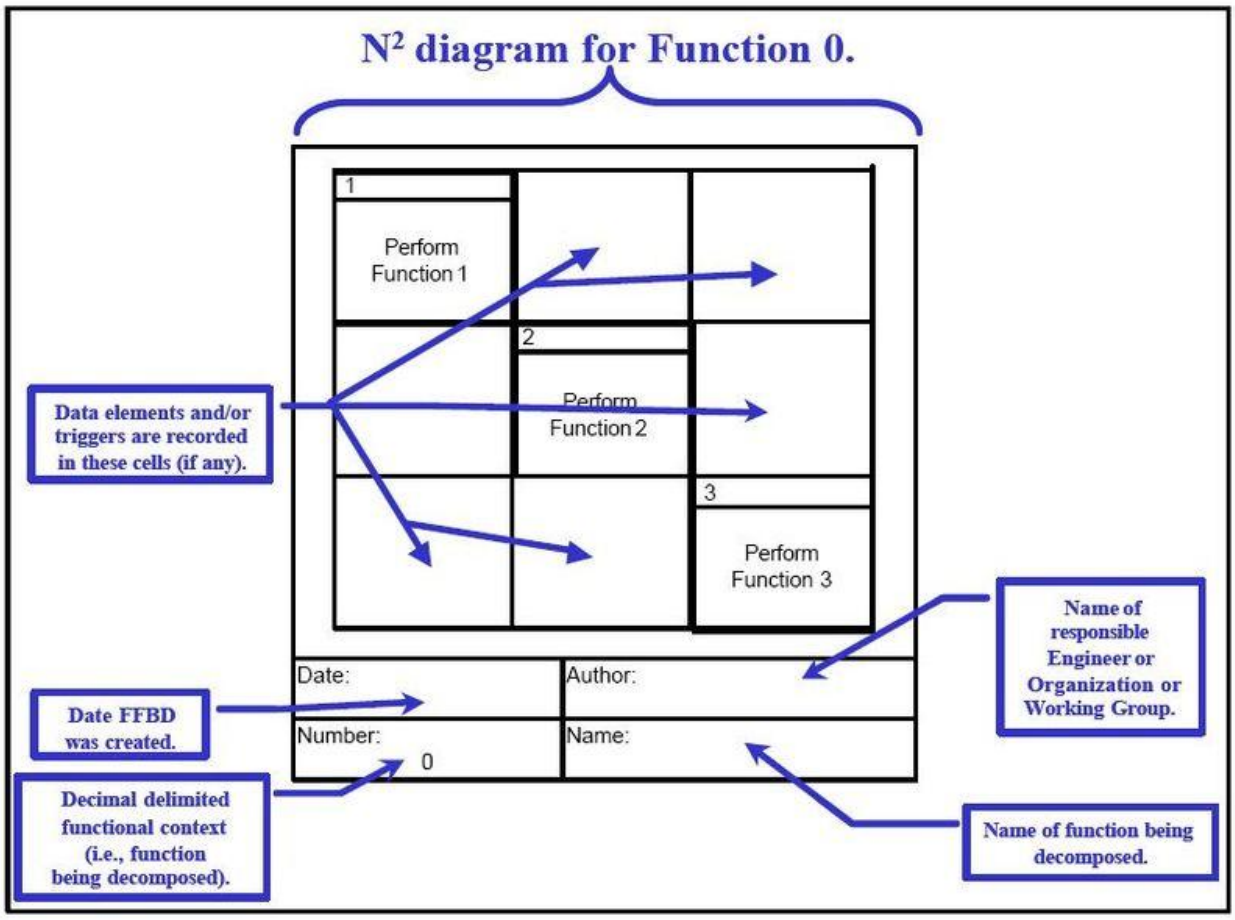


Figure 5 presents information in an N2 diagram, which complements the Functional flow block diagram. Notice that in this illustration, there are no data elements or triggers. The figure illustrates the context between functions at different levels of the model.

Review Questions

1. Define the Cartography?
2. Explain the Diagrams?
3. Explain the flow chart?
4. Explain the Square Diagram?

Discussion Questions

Discuss the history of cartography and its role in the development of modern cartography?

Chapter 2 - **Three Dimensional Diagrams**

Learning Objectives

- To define the Cartographic generalization.
- To explain the Cube Diagrams.
- To explain the Pictograms.

2.1 Cartographic generalization

Cartographic generalization is the method whereby information is selected and represented on a map in a way that adapts to the scale of the display medium of the map, not necessarily preserving all intricate geographical or other cartographic details. The cartographer is given license to adjust the content within their maps to create a suitable and useful map that conveys geospatial information, while striking the right balance between the map's purpose and actuality of the subject being mapped.

Well generalized maps are those that emphasize the most important map elements while still representing the world in the most faithful and recognizable way. The level of detail and importance in what is remaining on the map must outweigh the insignificance of items that were generalized, as to preserve the distinguishing characteristics of what makes the map useful and important.

2.1.1 **Methods**

Some cartographic generalization methods include the following:

2.1.1.1 **Selection**

Map generalization is designed to reduce the complexities of the real world by strategically reducing ancillary and unnecessary details. One way that geospatial data can be reduced is through the selection process. The cartographer can select and retain certain elements that he/she deems the most necessary or appropriate. In this method, the most important elements stand out while lesser elements are left out entirely. For example, a directional map between two points may have lesser and un-traveled roadways omitted as not to confuse the map-reader. The selection of the most direct and uncomplicated route between the two points is the most important data, and the cartographer may choose to emphasize this.

2.1.1.2 **Simplification**

Generalization is not a process that only removes and selects data, but also a process that simplifies it as well. Simplification is a technique where shapes of retained features are altered to enhance visibility and reduce complexity. Smaller scale maps have more simplified features than larger scale maps because they simply exhibit more area. An example of simplification is to scale and remove points along an area. Doing this to a mountain would reduce the detail in and around the mountain but would ideally not detract from the map reader interpreting the feature as such a mountain.

2.1.1.3 **Combination**

Simplification also takes on other roles when considering the role of combination. Overall data reduction techniques can also mean that in addition to generalizing elements of particular features, features can also be combined when their separation is irrelevant to the map focus. A mountain chain may be isolated into several smaller ridges and peaks with intermittent forest in the natural environment, but shown as a contiguous chain on the map, as determined by the scale. The map reader has to, again remember, that because of scale limitations combined elements are not concise depictions of natural or man-made features.

2.1.1.4 **Smoothing**

Smoothing is also a process that the map maker can employ to reduce the angularity of line work. Smoothing is yet another way of simplifying the map features, but involves several other characteristics of generalization that lead into feature displacement and locational shifting. The purpose of smoothing is to exhibit linework in a much less complicated and a less visually jarring way. An example of smoothing would be for a jagged roadway, cut through a mountain, to be smoothed out so that the angular turns and transitions appear much more fluid and natural.

2.1.1.5 **Enhancement**

Enhancement is also a method that can be employed by the cartographer to illuminate specific elements that aid in map reading. As many of the aforementioned generalizing methods focus on the reduction and omission of detail, the enhancement method concentrates on the addition of detail. Enhancement can be used to show the true character of the feature being represented and is often used by the cartographer to highlight specific details about his or her specific knowledge, that would otherwise be left out. An example includes enhancing the detail about specific river rapids so that the map reader may know the facets of traversing the most difficult sections beforehand. Enhancement can be a valuable tool in aiding the map reader to elements that carry significant weight to the map's intent.

2.1.2 **GIS and automated generalization**

As GIS gained prevalence in the late 20th century and the demand for producing maps automatically increased automated generalization became an important issue for National Mapping Agencies (NMAs) and other data providers. Thereby automated generalization is the automated extraction of data (becoming then information) regarding the purpose and scale. Different researchers invented conceptual models for automated generalization:

- Gruenreich model
- Brassel & Weibel model
- McMaster & Shea model

Besides these established models, different views on automated generalization have been established: the representation-oriented view and the process-oriented view. The first view focuses on the representation of data on different scales, which is related to the field of Multi-Representation Databases (MRDB). The latter view focuses on the process of generalization.

In the context of creating databases on different scales, additionally it can be distinguished between the ladder and the star-approach. The ladder-approach is a stepwise generalization, in which each derived

dataset is based on the other database of the next larger scale. The star-approach is the derived data on all scales is based on a single (large-scale) data base.

2.1.3 **Operators in automated generalization**

Automated generalization had always to compete with manual cartographers, therefore the manual generalization process was studied intensively. These studies resulted early in different generalization operators. By now there is no clear classification of operators available and it is doubtful if a comprehensive classification will evolve in future.

2.1.4 **The Baltimore Phenomenon**

The Baltimore Phenomenon is the tendency for a city to be omitted from maps due to space constraints while many smaller cities are included on the same map simply because the space is available to display them. This phenomenon gets its name from Baltimore, Maryland, which, despite its large population, is commonly omitted on maps of the United States because there is not enough space in the surrounding area of the map. Larger cities surrounding Baltimore take precedence. In contrast, much smaller cities in other geographic locations are included at the same scale because the level of competition for map space may not exist in that particular area.

2.1.5 **Competition for Limited Map Space**

During the design process of any map, either created manually or electronically, there will always be a finite amount of map space and an almost infinite amount of information that can be included in that space. Voids in the map will be present in rural areas where population is not very dense. This creates an easier decision-making process for the cartographer since most cities can be shown. There is plenty of space and therefore very little competition for that space by objects or points to be displayed. In contrast, densely populated areas create the constraint of working with a limited spatial area for both points representing cities and the labels of those cities being displayed. Baltimore is the largest city in Maryland, but due to its close proximity to Washington, D.C., and the necessity of labeling the oddly-shaped state of Maryland, it is omitted on maps in favor of other places that fit more easily in the map space available. Baltimore, despite its population, is omitted because of the necessary "competition for limited map space" in that geographic area.

2.1.6 **Baltimore in online mapping sites**

The Baltimore Phenomenon does not hold consistently true for every automated mapping site at every scale. Google Maps will display Baltimore once zoomed into the 7th zoom level. At the 6th zoom level, Baltimore is not displayed but cities such as Annapolis, Maryland, and Newton, Iowa, are displayed. Yahoo Maps display the major roads surrounding Baltimore at the 6th zoom level, but no city label appears until the 7th zoom level. Bing Maps display Baltimore beginning at the 5th zoom level, but other cities and surrounding details at this level are fairly sparse. OpenStreetMap is similar to Bing in that it displays Baltimore more readily than Google or Yahoo.

2.2 Cube Diagrams

In geometry, a **cube** is a three-dimensional solid object bounded by six square faces, facets or sides, with three meeting at each vertex.

The cube is the only **regular hexahedron** and is one of the five Platonic solids.

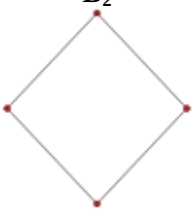
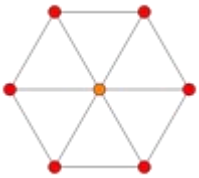
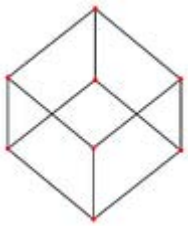
The cube is also a square parallelepiped, an equilateral cuboid and a right rhombohedron. It is a regular square prism in three orientations, and a trigonal trapezohedron in four orientations.

The cube is dual to the octahedron. It has cubical or octahedral symmetry.

2.2.1 **Orthogonal projections**

The *cube* has four special orthogonal projections, centered, on a vertex, edges, face and normal to its vertex figure. The first and third correspond to the A_2 and B_2 Coxeter planes.

Orthogonal projections

Centered by	**Face**	**Vertex**
	B_2	**A_2**
Coxeter planes		
Projective symmetry	[4]	[6]
Tilted views		

2.2.2 **Cartesian coordinates**

For a cube centered at the origin, with edges parallel to the axes and with an edge length of 2, the Cartesian coordinates of the vertices are

$$(\pm 1, \pm 1, \pm 1)$$

while the interior consists of all points (x_0, x_1, x_2) with $-1 < x_i < 1$.

2.2.3 **Equation in $\mathbf{R^3}$**

In analytic geometry, a cube's surface with center (x_0, y_0, z_0) and edge length of $2a$ is the locus of all points (x, y, z) such that

$$\lim_{n\to\infty} (x - x_0)^n + (y - y_0)^n + (z - z_0)^n - a^n = 0.$$

2.2.4 **Formulae**

For a cube of edge length a,

Surface area	$6a^2$
Volume	a^3
Face diagonal	$\sqrt{2}a$
Space diagonal	$\sqrt{3}a$
Radius of circumscribed sphere	$\frac{\sqrt{3}}{2}a$
Radius of sphere tangent to edges	$\frac{a}{\sqrt{2}}$
Radius of inscribed sphere	$\frac{a}{2}$
Angles between faces (in radians)	$\frac{\pi}{2}$

As the volume of a cube is the third power of its sides $a \times a \times a$, third powers are called *cubes*, by analogy with squares and second powers.

A cube has the largest volume among cuboids (rectangular boxes) with a given surface area. Also, a cube has the largest volume among cuboids with the same total linear size (length+width+height).

2.2.5 **Uniform colorings and symmetry**

The cube has three uniform colorings, named by the colors of the square faces around each vertex: 111, 112, 123.

The cube has three classes of symmetry, which can be represented by vertex-transitive coloring the faces. The highest octahedral symmetry O_h has all the faces the same color. The dihedral symmetry D_{4h} comes from the cube being a prism, with all four sides being the same color. The lowest symmetry D_{2h} is also a prismatic symmetry, with sides alternating colors, so there are three colors, paired by opposite sides. Each symmetry form has a different Wythoff symbol.

2.2.6 **Geometric relations**

A cube has eleven nets (one shown above): that is, there are eleven ways to flatten a hollow cube by cutting seven edges. To color the cube so that no two adjacent faces have the same color, one would need at least three colors.

The cube is the cell of the only regular tiling of three-dimensional Euclidean space. It is also unique among the Platonic solids in having faces with an even number of sides and, consequently, it is the only member of that group that is a zonohedron (every face has point symmetry).

The cube can be cut into six identical square pyramids. If these square pyramids are then attached to the faces of a second cube, a rhombic dodecahedron is obtained (with pairs of coplanar triangles combined into rhombic faces.)

2.2.7 **Other dimensions**

The analogue of a cube in four-dimensional Euclidean space has a special name—a tesseract or hypercube. More properly, a hypercube (or *n*-dimensional cube or simply *n*-cube) is the analogue of the cube in *n*-dimensional Euclidean space and a tesseract is the order-4 hypercube. A hypercube is also called a *measure polytope*.

There are analogues of the cube in lower dimensions too: a point in dimension 0, a segment in one dimension and a square in two dimensions.

2.2.8 **Related polyhedra**

The quotient of the cube by the antipodal map yields a projective polyhedron, the hemicube.

If the original cube has edge length 1, its dual polyhedron (an octahedron) has edge length $\sqrt{2}$.

The cube is a special case in various classes of general polyhedra:

Name	Equal edge-lengths?	Equal angles?	Right angles?
Cube	**Yes**	**Yes**	**Yes**
Rhombohedron	Yes	Yes	No
Cuboid	No	Yes	Yes
Parallelepiped	No	Yes	No
quadrilaterally faced hexahedron	No	No	No

The vertices of a cube can be grouped into two groups of four, each forming a regular tetrahedron; more generally this is referred to as a demicube. These two together form a regular compound, the stella octangula. The intersection of the two forms a regular octahedron. The symmetries of a regular tetrahedron correspond to those of a cube which map each tetrahedron to itself; the other symmetries of the cube map the two to each other.

One such regular tetrahedron has a volume of $\frac{1}{2}$ of that of the cube. The remaining space consists of four equal irregular tetrahedra with a volume of $\frac{1}{6}$ of that of the cube, each.

The rectified cube is the cuboctahedron. If smaller corners are cut off we get a polyhedron with six octagonal faces and eight triangular ones. In particular we can get regular octagons (truncated cube). The rhombicuboctahedron is obtained by cutting off both corners and edges to the correct amount.

A cube can be inscribed in a dodecahedron so that each vertex of the cube is a vertex of the dodecahedron and each edge is a diagonal of one of the dodecahedron's faces; taking all such cubes gives rise to the regular compound of five cubes.

If two opposite corners of a cube are truncated at the depth of the three vertices directly connected to them, an irregular octahedron is obtained. Eight of these irregular octahedra can be attached to the triangular faces of a regular octahedron to obtain the cuboctahedron.

2.2.9 **Combinatorial cubes**

A different kind of cube is the *cube graph*, which is the graph of vertices and edges of the geometrical cube. It is a special case of the hypercube graph.

An extension is the three dimensional *k*-ary Hamming graph, which for $k = 2$ is the cube graph. Graphs of this sort occur in the theory of parallel processing in computers.

2.3 Pictograms

A **pictogram**, also called a **pictogramme**, **pictograph**, or simply *picto*, and also an *is an ideogram that conveys its meaning through its pictorial resemblance to a physical object. Pictographs are often used in writing and graphic systems in which the characters are to a considerable extent pictorial in appearance*.

Pictography is a form of writing which uses representational, pictorial drawings, similar to cuneiform and, to some extent, hieroglyphic writing, which also uses drawings as phonetic letters or determinative rhymes. In certain modern use, pictographs participate to a formal language (e.g. Hazards pictograms).

2.3.1 **Historical**

Early written symbols were based on pictographs (pictures which resemble what they signify) and ideograms (symbols which represent ideas). Ancient Sumerian, Egyptian, and Chinese civilizations began to use such symbols over, developing them into logographic writing systems. Pictographs are still in use as the main medium of written communication in some non-literate cultures in Africa, The Americas, and Oceania. Pictographs are often used as simple, pictorial, representational symbols by most contemporary cultures.

Pictographs can be considered an art form, or can be considered a written language and are designated as such in Pre-Columbian art, Native American art, Ancient Mesopotamia and Painting in the Americas before Colonization. One example of many is the Rock art of the Chumash people, part of the Native American history of California. In 2011, UNESCO World Heritage adds to its list a new site "Petroglyph Complexes of the Mongolian Altai, Mongolia" to celebrate the importance of the pictographs engraved in rocks.

Some scientists in the field of neuropsychiatry and neuropsychology, such as Prof. Dr. Mario Christian Meyer, are studying the symbolic meaning of indigenous pictographs and petroglyphs, aiming to create new ways of communication between native people and modern scientists to safeguard and valorize their cultural diversity.

2.3.2 **Modern uses**

An early modern example of the extensive use of pictographs may be seen in the map in the London suburban timetables of the London and North Eastern Railway, 1936-1947, designed by George Dow, in which a variety of pictographs were used to indicate facilities available at or near each station. Pictographs remain in common use today, serving as a pictorial, representational signs, instructions, or

statistical diagrams. Because of their graphical nature and fairly realistic style, they are widely used to indicate public toilets, or places such as airports and train stations.

Pictographic writing as a modernist poetic technique is credited to Ezra Pound, though French surrealists accurately credit the Pacific Northwest American Indians of Alaska who introduced writing, via totem poles, to North America.

Contemporary artist Xu Bing created Book from the Ground, a universal language made up of pictographs collected from around the world. A Book from the Ground chat program has been exhibited in museums and galleries internationally.

Pictographs are used in many areas of modern life for commodity purposes, often as a formal language.

2.3.3 **Standardization**

Pictographs can often transcend languages in that they can communicate to speakers of a number of tongues and language families equally effectively, even if the languages and cultures are completely different. This is why road signs and similar pictographic material are often applied as global standards expected to be understood by nearly all.

A standard set of pictographs was defined in the international standard *ISO 7001: Public Information Symbols.* Another common set of pictographs are the laundry symbols used on clothing tags and the chemical hazard symbols as standardized by the GHS system.

Pictographs have been popularized in use on the internet and in softwares, better known as "icons" displayed on a computer screen in order to help the user navigate a computer system or mobile device.

Review Questions

1. Define the Cartographic generalization?
2. Explain the Cube Diagrams?
3. Explain the Pictograms?

Discussion Questions

Discuss the Cartographic generalization methods?

Chapter 3- Distributional Maps

Learning Objectives

- To define the map.
- To explain the Isopleth Maps.
- To explain the Choropleth Maps.
- To describe the Dot Maps.

3.1 Map

A **map** is a visual representation of an area – symbolic depiction highlighting relationships between elements of that space such as objects, regions, and themes.

Many maps are static two-dimensional, geometrically accurate (or approximately accurate) representations of three-dimensional space, while others are dynamic or interactive, even three-dimensional. Although most commonly used to depict geography, maps may represent any space, real or imagined, without regard to context or scale; e.g. brain mapping, DNA mapping and extraterrestrial mapping.

Although the earliest maps known are of the heavens, geographic maps of territory have a very long tradition and exist from ancient times. The word "map" comes from the medieval Latin *Mappa mundi*, wherein *mappa* meant napkin or cloth and *mundi* the world. Thus, "map" became the shortened term referring to a 2 dimensional representation of the surface of the world.

3.1.1 Geographic maps

Cartography or *map-making* is the study and practice of crafting representations of the Earth upon a flat surface, and one who makes maps is called a cartographer.

Road maps are perhaps the most widely used maps today, and form a subset of navigational maps, which also include aeronautical and nautical charts, railroad network maps, and hiking and bicycling maps. In terms of quantity, the largest number of drawn map sheets is probably made up by local surveys, carried out by municipalities, utilities, tax assessors, emergency service providers, and other local agencies. Many national surveying projects have been carried out by the military, such as the British Ordnance Survey (now a civilian government agency internationally renowned for its comprehensively detailed work).

In addition to location information maps may also be used to portray contour lines indicating constant values of elevation, temperature, rainfall, etc.

3.1.2 Orientation of maps

The orientation of a map is the relationship between the directions on the map and the corresponding compass directions in reality. The word "orient" is derived from Latin *oriens*, meaning East. In the Middle Ages many maps, including the T and O maps, were drawn with East at the top (meaning that the direction "up" on the map corresponds to the East on the compass). Today, the most common – but far

from universal – cartographic convention is that North is at the top of a map. Several kinds of maps are often traditionally not oriented with North at the top:

- Maps from non-Western traditions are oriented a variety of ways. Old maps of Edo show the Japanese imperial palace as the "top", but also at the center, of the map. Labels on the map are oriented in such a way that you cannot read them properly unless you put the imperial palace above your head.
- Medieval European T and O maps such as the Hereford Mappa Mundi were centered on Jerusalem with East at the top. Indeed, prior to the reintroduction of Ptolemy's *Geography* to Europe around 1400, there was no single convention in the West. Portolan charts, for example, are oriented to the shores they describe.
- Maps of cities bordering a sea are often conventionally oriented with the sea at the top.
- Route and channel maps have traditionally been oriented to the road or waterway they describe.
- Polar maps of the Arctic or Antarctic regions are conventionally centered on the pole; the direction North would be towards or away from the center of the map, respectively. Typical maps of the Arctic have 0° meridian towards the bottom of the page; maps of the Antarctic have the 0° meridian towards the top of the page.
- Reversed maps, also known as Upside-Down maps or South-Up maps, reverse the "North is up" convention and have South at the top.
- Buckminster Fuller's Dymaxion maps are based on a projection of the Earth's sphere onto an icosahedron. The resulting triangular pieces may be arranged in any order or orientation.
- Modern digital GIS maps such as ArcMap typical project north at the top of the map, but use math degrees (0 is east, degrees increase counter-clockwise), rather than compass degrees (0 is north, degrees increase clockwise) for orientation of transects. Compass decimal degrees can be converted to math degrees by subtracting them from 450; if the answer is greater than 360, subtract 360.

3.1.3 Scale and accuracy

Many, but not all, maps are drawn to a scale, expressed as a ratio such as 1:10,000, meaning that 1 of any unit of measurement on the map corresponds exactly, or approximately, to 10,000 of that same unit on the ground. The scale statement may be taken as exact when the region mapped is small enough for the curvature of the Earth to be neglected, for example in a town planner's city map. Over larger regions where the curvature cannot be ignored we must use map projections from the curved surface of the Earth (sphere or ellipsoid) to the plane. The impossibility of flattening the sphere to the plane implies that no map projection can have constant scale: on most projections the best we can achieve is an accurate scale on one or two lines (not necessarily straight) on the projection. Thus for map projections we must introduce the concept of point scale, which is a function of position, and strive to keep its variation within narrow bounds. Although the scale statement is nominal it is usually accurate enough for all but the most precise of measurements.

Large scale maps, say 1:10,000, cover relatively small regions in great detail and small scale maps, say 1:10,000,000, cover large regions such as nations, continents and the whole globe. The large/small terminology arose from the practice of writing scales as numerical fractions: 1/10,000 is larger than 1/10,000,000. There is no exact dividing line between large and small but 1/100,000 might well be considered as a medium scale. Examples of large scale maps are the 1:25,000 maps produced for hikers; on the other hand maps intended for motorists at 1:250,000 or 1:1,000,000 are small scale.

It is important to recognize that even the most accurate maps sacrifice a certain amount of accuracy in scale to deliver a greater visual usefulness to its user. For example, the width of roads and small streams

are exaggerated when they are too narrow to be shown on the map at true scale; that is, on a printed map they would be narrower than could be perceived by the naked eye. The same applies to computer maps where the smallest unit is the pixel. A narrow stream say must be shown to have the width of a pixel even if at the map scale it would be a small fraction of the pixel width.

Some maps, called cartograms, have the scale deliberately distorted to reflect information other than land area or distance. For example, this map (at the right) of Europe has been distorted to show population distribution, while the rough shape of the continent is still discernible.

Another example of distorted scale is the famous London Underground map. The basic geographical structure is respected but the tube lines (and the River Thames) are smoothed to clarify the relationships between stations. Near the center of the map stations are spaced out more than near the edges of the map.

Further inaccuracies may be deliberate. For example, cartographers may simply omit military installations or remove features solely in order to enhance the clarity of the map. For example, a road map may not show railroads, smaller waterways or other prominent non-road objects, and even if it does, it may show them less clearly (e.g. dashed or dotted lines/outlines) than the main roads. Known as decluttering, the practice makes the subject matter that the user is interested in easier to read, usually without sacrificing overall accuracy. Software-based maps often allow the user to toggle decluttering between ON, OFF and AUTO as needed. In AUTO the degree of decluttering is adjusted as the user changes the scale being displayed.

3.1.4 Map types and projections

The purpose of the physical is to show features of geography such as mountains, soil type or land use including infrastructure such as roads, railroads and buildings. Topographic maps show elevations and relief with contour lines or shading. Geological maps show not only the physical surface, but characteristics of the underlying rock, fault lines, and subsurface structures.

Maps that depict the surface of the Earth also use a projection, a way of translating the three-dimensional real surface of the geoid to a two-dimensional picture. Perhaps the best-known world-map projection is the Mercator projection, originally designed as a form of nautical chart.

Aero plane pilots use aeronautical charts based on a Lambert conformal conic projection, in which a cone is laid over the section of the earth to be mapped. The cone intersects the sphere (the earth) at one or two parallels which are chosen as standard lines. This allows the pilots to plot a great-circle route approximation on a flat, two-dimensional chart.

- Azimuthal or Gnomonic map projections are often used in planning air routes due to their ability to represent great circles as straight lines.
- General Richard Edes Harrison produced a striking series of maps during and after World War II for Fortune magazine. These used "bird's eye" projections to emphasize globally strategic "fronts" in the air age, pointing out proximities and barriers not apparent on a conventional rectangular projection of the world.

3.1.5 Electronic maps

From the last quarter of the 20th century, the indispensable tool of the cartographer has been the computer. Much of cartography, especially at the data-gathering survey level, has been subsumed by

Geographic Information Systems (GIS). The functionality of maps has been greatly advanced by technology simplifying the superimposition of spatially located variables onto existing geographical maps. Having local information such as rainfall level, distribution of wildlife, or demographic data integrated within the map allows more efficient analysis and better decision making. In the pre-electronic age such superimposition of data led Dr. John Snow to identify the location of an outbreak of cholera. Today, it is used by agencies of the human kind, as diverse as wildlife conservationists and militaries around the world.

Even when GIS is not involved, most cartographers now use a variety of computer graphics programs to generate new maps.

Interactive, computerized maps are commercially available, allowing users to *zoom in* or *zoom out* (respectively meaning to increase or decrease the scale), sometimes by replacing one map with another of different scale, centered where possible on the same point. In-car global navigation satellite systems are computerized maps with route-planning and advice facilities which monitor the user's position with the help of satellites. From the computer scientist's point of view, zooming in entails one or a combination of:

1. Replacing the map by a more detailed one
2. Enlarging the same map without enlarging the pixels, hence showing more detail by removing less information compared to the less detailed version
3. Enlarging the same map with the pixels enlarged (replaced by rectangles of pixels); no additional detail is shown, but, depending on the quality of one's vision, possibly more detail can be seen; if a computer display does not show adjacent pixels really separate, but overlapping instead (this does not apply for an LCD, but may apply for a cathode ray tube), then replacing a pixel by a rectangle of pixels does show more detail. A variation of this method is an interpolation.

For example:

- Typically (2) applies to a Portable Document Format (PDF) file or other format based on vector graphics. The increase in detail is, of course, limited to the information contained in the file: enlargement of a curve may eventually result in a series of standard geometric figures such as straight lines, arcs of circles or splines.
- (2) May apply to text and (3) to the outline of a map feature such as a forest or building.
- (1) May apply to the text as needed (displaying labels for more features), while (2) applies to the rest of the image. Text is not necessarily enlarged when zooming in. Similarly, a road represented by a double line may or may not become wider when one zooms in.
- The map may also have layers which are partly raster graphics and partly vector graphics. For a single raster graphics image (2) applies until the pixels in the image file correspond to the pixels of the display, thereafter (3) applies.

3.1.6 Conventional signs

The various features shown on a map are represented by conventional signs or symbols. For example, colors can be used to indicate a classification of roads. Those signs are usually explained in the margin of the map, or on a separately published characteristic sheet.

Some cartographers prefer to make the map cover practically the entire screen or sheet of paper, leaving no room "outside" the map for information about the map as a whole. These cartographers typically place such information in an otherwise "blank" region "inside" the map -- cartouche, map legend, title, compass rose, bar scale, etc. In particular, some maps contain smaller "sub-maps" in otherwise blank regions—

often one at a much smaller scale showing the whole globe and where the whole map fits on that globe, and a few showing "regions of interest" at a larger scale in order to show details that wouldn't otherwise fit. Occasionally sub-maps use the same scale as the large map—a few maps of the contiguous United States include a sub-map to the same scale for each of the two non-contiguous states.

3.1.7 Labeling

To communicate spatial information effectively, features such as rivers, lakes, and cities need to be labeled. Over centuries cartographers have developed the art of placing names on even the densest of maps. Text placement or name placement can get mathematically very complex as the number of labels and map density increases. Therefore, text placement is time-consuming and labor-intensive, so cartographers and GIS users have developed automatic label placement to ease this process.

3.1.8 Non-geographical spatial maps

Maps exist in the solar system, and other cosmological features such as star maps. In addition to maps of other bodies such as the Moon and other planets are technically not *geo*graphical maps.

3.1.9 Non spatial maps

Diagrams such as schematic diagrams and Gantt charts and treemaps display logical relationships between items, and do not display spatial relationships at all.

Some maps, for example the London Underground map, are topological maps. Topological in nature, the distances are completely unimportant; only the connectivity is significant.

3.1.10 General-purpose maps

General-purpose maps provide many types of information on one map. Most atlas maps, wall maps, and road maps fall into this category. The following are some features that might be shown on a general-purpose map: bodies of water, roads, railway lines, parks, elevations, towns and cities, political boundaries, latitude and longitude, national and provincial parks. These maps give a broad understanding of location and features of an area. You can gain an understanding of the type of landscape, the location of urban places, and the location of major transportation routes all at once.

3.2 Isopleth Maps

A **contour line** (also **isoline**, **isopleth**, or **isarithm**) of a function of two variables is a curve along which the function has a constant value. In map-making, a contour line (often just called a "contour") joins points of equal elevation (height) above a given level, such as mean sea level. A **contour map** is a map illustrated with contour lines, for example a topographic map, which thus shows valleys and hills, and the steepness of slopes. The **contour interval** of a contour map is the difference in elevation between successive contour lines.

More generally, a contour line for a function of two variables is a curve connecting points where the function has the same particular value. The gradient of the function is always perpendicular to the contour lines. When the lines are close together the magnitude of the gradient is large: the variation is steep. A level set is a generalization of a contour line for functions of any number of variables.

Contour lines are curved or straight lines on a map describing the intersection of a real or hypothetical surface with one or more horizontal plane. The configuration of these contours allows map readers to infer relative gradient of a parameter and estimate that parameter at specific places. Contour lines may be either traced on a visible three-dimensional model of the surface, as when a photogrammetrist viewing a stereo-model plots elevation contours, or interpolated from estimated surface elevations, as when a computer program threads contours through a network of observation points of area centroids. In the latter case, the method of interpolation affects the reliability of individual isolines and their portrayal of slope, pits and peaks.

3.2.1 **Types**

Contour lines are often given specific names beginning "iso-" , according to the nature of the variable being mapped, although in many usages the phrase "contour line" is most commonly used. Specific names are most common in meteorology, where multiple maps with different variables may be viewed simultaneously. The prefix "iso-" can be replaced with "isallo-" to specify a contour line connecting points where a variable changes at the same *rate* during a given time period.

The words *isoline* and *isarithm* are general terms covering all types of contour line. The word *isogram* was proposed by Francis Galton in 1889 as a convenient generic designation for lines indicating equality of some physical condition or quantity; but it commonly refers to a word without a repeated letter.

An **isogon** is a contour line for a variable which measures direction. In meteorology and in geomagnetics, the term *isogon* has specific meanings which are described below. An **insulin** is a line joining points with equal slope. In population dynamics and in geomagnetics, the terms *isocline* and *isoclinic line* have specific meanings which are described below.

3.2.1.1 **Equidistants (isodistances)**

Equidistant is a line of equal distance from a given point, line, polyline.

3.2.1.2 **Isopleths**

In geography, the word *isopleth* (from πλῆθος or *plethos*, meaning 'quantity') is used for contour lines that depict a variable which cannot be measured at a point, but which instead must be calculated from data collected over an area. An example is population density, which can be calculated by dividing the population of a census district by the surface area of that district. Each calculated value is presumed to be the value of the variable at the center of the area, and isopleths can then be drawn by a process of interpolation. The idea of an isopleth map can be compared with that of a choropleth map.

In meteorology, the word *isopleth* is used for any type of the contour line.

3.2.1.3 **Meteorology**

Meteorological contour lines are based on generalization from the point data received from weather stations. Weather stations are seldom exactly positioned at a contour line (when they are, this indicates a measurement precisely equal to the value of the contour). Instead, lines are drawn to best approximate the locations of exact values, based on the scattered information points available.

Meteorological contour maps may present collected data such as actual air pressure at a given time, or generalized data such as the average pressure over a period of time, or forecast data such as predicted air pressure at some point in the future

Thermodynamic diagrams use multiple overlapping contour sets (including isobars and isotherms) to present a picture the major thermodynamic factors in a weather system.

3.2.1.4 ***Barometric pressure***

An **isobar** is a line of equal or constant pressure on a graph, plot, or map; an isopleth or contour line of pressure. More accurately, isobars are lines drawn on a map joining places of equal average atmospheric pressure reduced to sea level for a specified period of time. In meteorology, the barometric pressures shown are reduced to sea level, not the surface pressures at the map locations. The distribution of isobars is closely related to the magnitude and direction of the wind field, and can be used to predict future weather patterns. Isobars are commonly used in television weather reporting.

An **isostere** is a line of constant atmospheric density. An **isoheight** or **isohypse** is a line of constant geopotential height on a constant pressure surface chart.

3.2.1.5 ***Temperature and related subjects***

An **isotherm** (from θέρμη or *thermē*, meaning 'heat') is a line that connects points on a map that have the same temperature. Therefore, all points through which an isotherm passes have the same or equal temperatures at the time indicated. An isotherm at 0°C is called the freezing level.

An **isogeotherm** is a line of equal mean annual temperature. An **isocheim** is a line of equal mean winter temperature, and an **isothere** is a line of equal mean summer temperature.

An **isohel** is a line of equal or constant solar radiation.

3.2.1.6 ***Precipitation and air moisture***

An **isohyet** or **isohyetal line** is a line joining points of equal precipitation on a map. A map with isohyets is called an **isohyetal map**.

An **isohume** is a line of constant relative humidity, while a **isodrosotherm** (from δρόσος or *drosos*, meaning 'dew', and or *therme*, meaning 'heat') is a line of equal or constant dew point.

An **isoneph** is a line indicating equal cloud cover.

An **isochalaz** is a line of constant frequency of hail storms, and an **isobront** is a line drawn through geographical points at which a given phase of thunderstorm activity occurred simultaneously.

Snow cover is frequently shown as a contour-line map.

3.2.1.7 ***Wind***

An **isotach** is a line joining points with constant wind speed. In meteorology, the term **isogon** refers to a line of constant wind direction.

3.2.1.8 ***Freeze and thaw***

An **isopectic** line denotes equal dates of ice formation each winter, and an **isotac** denotes equal dates of thawing.

3.2.2 **Physical geography and oceanography**

3.2.2.1 ***Elevation and depth***

Contours are one of several common methods used to denote elevation or altitude and depth on maps. From these contours, a sense of the general terrain can be determined. They are used at a variety of scales, from large-scale engineering drawings and architectural plans, through topographic maps up to continental-scale maps.

"Contour line" is the most common usage in map-making, but isobath for underwater depths on bathymetric maps and **isohypse** for elevations are also used. The process of drawing isohypse contour lines on a map is called *isopletion*.

In map-making, the **contour interval** is the elevation difference between adjacent contour lines. The contour interval should be the same over a single map. When calculated as a ratio against the map scale, a sense of the hilliness of the terrain can be derived.

3.2.2.2 ***Magnetism***

In the study of the Earth's magnetic field, the term **isogon** or **isogonic line** refers to a line of constant magnetic declination, the variation of magnetic north from geographic north. An **agonic line** is drawn through points of zero magnetic declination.

An **isoclinic line** connects points of equal magnetic dip, and an **aclinic line** is the isoclinic line of magnetic dip zero.

An **isodynamic line** connects points with the same intensity of the magnetic force.

3.2.3 ***Oceanography***

Besides the ocean depth, oceanographers use contour to describe diffuse variable phenomena much as meteorologists do with atmospheric phenomena. In particular, **isobathytherms** are lines showing depths of water with equal temperature, **isohalines** show lines of equal ocean salinity, and **Isopycnals** are surfaces of equal water density.

3.2.4 **Geology**

Various geological data are rendered as contour maps in structural geology, sedimentology, stratigraphy and economic geology. Contour maps are used to show the below ground surface of geologic strata, fault surfaces (especially low angle thrust faults) and unconformities. Isopach maps use **isopachs** (lines of equal thickness) to illustrate variations in thickness of geologic units.

3.2.5 **Environmental science**

In discussing pollution, density maps can be very useful in indicating sources and areas of greatest contamination. Contour maps are especially useful for diffuse forms or scales of pollution. Acid precipitation is indicated on maps with **isoplats**. Some of the most widespread applications of environmental science contour maps involve mapping of environmental noise (where lines of equal sound pressure level are denoted **isobels**), air pollution, soil contamination, thermal pollution and groundwater contamination. By contour planting and contour ploughing, the rate of water runoff and thus soil erosion can be substantially reduced; this is especially important in riparian zones.

3.2.6 **Ecology**

An **isoflor** is an isopleth contour connecting areas of comparable biological diversity. Usually, the variable is the number of species of a given genus or family that occurs in a region. Isoflor maps are thus used to show distribution patterns and trends such as centers of diversity.

3.2.7 **Social sciences**

In economics, contour lines can be used to describe features which vary quantitatively over space. An isochrone shows lines of equivalent drive time or travel time to a given location and is used in the generation of isochrone maps. An **isotim** shows equivalent transport costs from the source of a raw material, and an **isodapane** shows equivalent cost of travel time.

Indifference curves are used to show bundles of goods to which a person would assign equal utility. In political science an analogous method is used in understanding coalitions (for example the diagram in Laver and Shepsle's work).

In population dynamics, **isocline** refers to the set of population sizes at which the rate of change, or partial derivative, for one population in a pair of interacting populations is zero.

Isolines can also be used to delineate qualitative differences. An **isogloss**, for example, is used in mapping the geographic spread of linguistic features.

Contour lines are also used in non-geographic charts in economics. An **isoquant** is a line of equal production quantity, and an **isocost** shows equal production costs.

3.2.9 **Thermodynamics, engineering, and other sciences**

Various types of graphs in thermodynamics, engineering, and other sciences use isobars (constant pressure), isotherms (constant temperature), isochors (constant specific volume), or other types of isolines, even though these graphs are usually not related to maps. Such isolines are useful for representing more than two dimensions (or quantities) on two-dimensional graphs. Common examples in thermodynamics are some types of phase diagrams.

Isoclines are used to solve ordinary differential equations.

In interpreting radar images, an **isodop** is a line of equal Doppler velocity, and an **isoecho** is a line of equal radar reflectivity.

3.2.10 **Other phenomena**

- isochasm: aurora equal occurrence
- isochor: volume
- isodose: radiation intensity
- isophene: biological events occurring with coincidence such as plants flowering
- isophote: illuminance

3.2.11 **History**

The idea of lines that join points of equal value was rediscovered several times. In 1701, Edmond Halley used such lines (isogons) on a chart of magnetic variation. The Dutch engineer Nicholas Cruquius drew the bed of the river Merwede with lines of equal depth (isobaths) at intervals of 1 fathom in 1727, and Philippe Buache used them at 10-fathom intervals on a chart of the English Channel that was prepared in 1737 and published in 1752. Such lines were used to describe a land surface (contour lines) in a map of the Duchy of Modena and Reggio by Domenico Vandelli in 1746, and they were studied theoretically by Ducarla in 1771, and Charles Hutton used them when calculating the volume of a hill in 1777. In 1791, a map of France by J. L. Dupain-Triel used contour lines at 20-metre intervals, hachures, spot-heights and a vertical section. In 1801, the chief of the Corps of Engineers, Haxo, used contour lines at the larger scale of 1:500 on a plan of his projects for Rocca d'Aufo.

By around 1843, when the Ordnance Survey started to regularly record contour lines in Great Britain and Ireland, they were already in general use in European countries. Isobaths were not routinely used on nautical charts until those of Russia from 1834, and those of Britain from 1838.

When maps with contour lines became common, the idea spread to other applications. Perhaps the latest to develop are air quality and noise pollution contour maps, which first appeared in the US, in approximately 1970, largely as a result of national legislation requiring spatial delineation of these parameters. In 2007, Pictometry International was the first to allow users to dynamically generate elevation contour lines to be laid over oblique images.

3.2.12 **Technical construction factors**

To maximize readability of contour maps, there are several design choices available to the map creator, principally line weight, line color, line type and method of numerical marking.

Line weight is simply the darkness or thickness of the line used. This choice is made based upon the least intrusive form of contours that enable the reader to decipher the background information in the map itself. If there is little or no content on the base map, the contour lines may be drawn with relatively heavy thickness. Also, for many forms of contours such as topographic maps, it is common to vary the line weight and/or color, so that a different line characteristic occurs for certain numerical values. For example, in the topographic map above, the even hundred foot elevations are shown in a different weight from the twenty foot intervals.

Line color is the choice of any number of pigments that suit the display. Sometimes a sheen or gloss is used as well as color to set the contour lines apart from the base map. Line colour can be varied to show other information.

Line type refers to whether the basic contour line is solid, dashed, dotted or broken in some other pattern to create the desired effect. Dotted or dashed lines are often used when the underlying base map conveys

very important (or difficult to read) information. Broken line types are used when the location of the contour line is inferred.

Numerical marking is the manner of denoting the arithmetical values of contour lines. This can be done by placing numbers along some of the contour lines, typically using interpolation for intervening lines. Alternatively a map key can be produced associating the contours with their values.

If the contour lines are not numerically labeled and adjacent lines have the same style (with the same weight, color and type), then the direction of the gradient cannot be determined from the contour lines alone. However if the contour lines cycle through three or more styles, then the direction of the gradient can be determined from the lines. The orientation of the numerical text labels is often used to indicate the direction of the slope.

3.2.13 **Plan views versus profile view**

Most commonly contour lines are drawn in plan view, or as an observer in space would view the Earth's surface: ordinary map form. However, some parameters can often be displayed in profile view showing a vertical profile of the parameter mapped. Some of the most common parameters mapped in the profile are air pollutant concentrations and sound levels. In each of those cases it may be important to analyze (air pollutant concentrations or sound levels) at varying heights so as to determine the air quality or noise health effects on people at different elevations, for example, living on different floor levels of an urban apartment. In actuality, both plan and profile view contour maps are used in air pollution and noise pollution studies.

3.2.14 **Labeling contour maps**

Labels are a critical component of elevation maps. A properly labeled contour map helps the reader to quickly interpret the shape of the terrain. If numbers are placed close to each other, it means that the terrain is steep. Labels should be placed along a slightly curved line "pointing" to the summit or nadir, from several directions if possible, making the visual identification of the summit or nadir easy. Contour labels can be oriented so a reader is facing uphill when reading the label.

Manual labeling of contour maps is a time-consuming process, however, there are a few software systems that can do the job automatically and in accordance with cartographic conventions, called automatic label placement.

3.3 Choropleth Maps

A **choropleth map**, is a thematic map in which areas are shaded or patterned in proportion to the measurement of the statistical variable being displayed on the map, such as population density or per-capita income.

The choropleth map provides an easy way to visualize how a measurement varies across a geographic area or it shows the level of variability within a region.

A special type of choropleth map is a prism map, a three-dimensional map in which a given region's height on the map is proportional to the statistical variable's value for that region.

3.3.1 Overview

The earliest known choropleth map was created in 1826 by Baron Pierre Charles Dupin. The term "choroplethe map" was introduced 1938 by the geographer John Kirtland Wright in "Problems in Population Mapping".

Choropleth maps are based on statistical data aggregated over previously defined regions (e.g., counties), in contrast to area-class and isarithmic maps, in which region boundaries are defined by data patterns. Thus, where defined regions are important to a discussion, as in an election map divided by electoral regions, choropleths are preferred.

Where real-world patterns may not conform to the regions discussed, issues such as the ecological fallacy and the modifiable areal unit problem (MAUP) can lead to major misinterpretations, and other techniques are preferable. Choropleth maps are frequently used in inappropriate applications due to the abundance of choropleth data and the ease of design using Geographic Information Systems.

The dasymetric technique can be thought of as a compromise approach in many situations. Broadly speaking choropleths represent two types of data: Spatially Extensive or Spatially Intensive.

- Spatially Extensive data are things like populations. The population of the UK might be 60 million, but it would not be accurate to arbitrarily cut the UK into two halves of equal area and say that the population of each half of the UK is 30 million.
- Spatially Intensive data are things like rates, densities and proportions, which can be thought of conceptually as field data that is averaged over an area. Though the UK's 60 million inhabitants occupy an area of about 240,000 km^2, and the population density is therefore about 250/km^2, arbitrary halves of equal area would not also both have the same population density.

Another common error in choropleths is the use of raw data values to represent magnitude rather than normalized values to produce a map of densities. This is problematic because the eye naturally integrates over areas of the same color, giving undue prominence to larger polygons of moderate magnitude and minimizing the significance of smaller polygons with high magnitudes. Compare the circled features in the maps at right.

3.3.2 Color progression

When mapping quantitative data, a specific color progression should be used to depict the data properly. There are several different types of color progressions used by cartographers. The following are described in detail in Robinson et al. (1995)

Single hue progression

Single-hue progressions fade from a dark shade of the chosen color to a very light or white shade of relatively the same hue. This is a common method used to map magnitude. The dark hue represents the greatest number in the data set and the lightest shade representing the least number.

Two variables may be shown through the use of two overprinted single color scales. The hues typically used are from red to white for the first data set and blue to white for the second, they are then overprinted to produce varying hues. These types of maps show the magnitude of the values in relation to each other.

Bipolar color progression

Bipolar progressions are normally used with two opposite hues to show a change in value from negative to positive or on either side of some either central tendency, such as the mean of the variable being mapped or other significant value like room temperature. For example a typical progression when mapping temperatures is from dark blue (for cold) to dark red (for hot) with white in the middle. When one extreme can be considered better than the other (as in this map of life expectancy) then it is common to denote the poor alternative with shades of red, and the best alternative with green.

Complementary hue progressions are a type of bi-polar progression. This can be done with any of the complementary colors and will fade from each of the darker end point hues into a gray shade representing the middle. An example would be using blue and yellow as the two end points.

Blended hue color progression

Blended hue progressions use related hues to blend together the two end point hues. This type of color progression is typically used to show elevation changes. For example from yellow through orange to brown.

Partial spectral color progression

Partial spectral hue progressions are used to map mixtures of two distinct sets of data. This type of hue progression will blend two adjacent opponent hues and show the magnitude of the mixing data classes.

Full-spectral color progression

Full spectral progression contains hues from blue through red. This is common on relief maps and modern weather maps. This type of progression is not recommended under other circumstances because certain color connotations can confuse the map user.

Value progression

Value progression maps are monochromatic. Although any color may be used, the archetype is from black to white with intervening shades of gray that represent magnitude. According to Robinson *et al.*

(1995). This is the best way to portray a magnitude message to the map audience. It is clearly understood by the user and easy to produce in print.

3.3.3 Usability

When using any of these methods there are two important principles: first is that darker colors are perceived as being higher in magnitude and second is that while there are millions of color variations the human eye is limited to how many colors it can easily distinguish. Generally five to seven color categories is recommended. The map user should be able to easily identify the implied magnitude of the hue and match it with the legend.

Additional considerations include color blindness and various reproduction techniques. For example, the red–green bi-polar progression described in the section above is likely to cause problems for dichromats. A related issue is that color scales which rely primarily on hue with insufficient variation in saturation or intensity may be compromised if reproduced with a black and white device; if a map is legible in black and white, then a prospective user's perception of color is irrelevant.

Color can greatly enhance the communication between the cartographer and their audience but poor color choices can result in a map that is neither effective nor appealing to the map user; sometimes simpler is better.

3.4 **Dot Maps**

A dot density map is as a map type that uses a dot symbol to show the presence of a feature or phenomenon. Dot maps rely on a visual scatter to show spatial pattern.

3.4.1 Types of dot maps

3.4.1.1 One-to-one

In a one-to-one dot map, each dot represents one single recording of a phenomenon. Because the location of the dot corresponds to only one piece of data, care must be taken to ensure that the dot is represented in its correct spatial location. Inaccuracies in the location of the dot can misrepresent the data being mapped. Various methods exist for determining the exact spatial location of a single point, including geocoding.

3.4.1.2 One-to-many

In a one-to-many, or dot-density map, each dot on the map represents more than one of the phenomena being mapped. The number of data represented by each dot is determined by the map author and may be the product of data availability. Some data, such as the addresses of cancer patients, may not be available for mapping due to restrictions on access to individuals' medical records.

In one-to-many dot distribution maps, the reader must be careful not to interpret the dots as actual locations, as the dots represent aggregate data and are often arbitrarily placed on a map. Methods of dot placement include by areal unit centroid, random disbursement, and uniform (evenly spaced) placement, among others.

3.4.2 Historical examples of dot distribution maps

3.4.2.1 *Carte philosophique figurant la population de la France*

The first dot distribution map was created by a Franciscan monk, Armand Joseph Frère de Montizon (1788 - ????). It is a relatively simple map of population by département (administrative district) in France and is one of the first known examples of a demographic map of the country. Each dot represents 10,000 individuals. The dots are spaced in even rows, the distance between which determined by the population of the department. A table in the map lists the departments by name, population, and prefectural city. The departments were numbered on the map to correspond to the table. The regular spacing of the dots in the map produces a visual display of population density, as higher population levels within an administrative border exhibit a closer, denser pattern of dots. Since the dots are evenly spaced, it is evident that they do not represent the actual locations of where people live within a department. This is an example of an ecological fallacy, where a value for an area generalizes all within that area to exhibit that value.

Although Montizon's map was the first thematic dot map published, it did not garner the author fame. Instead, his innovation had no effect on practice for nearly 30 years until the dot distribution map was "reinvented" for map by a Swedish Army officer, published in 1859. This map was authored by Thure Alexander von Mentzer and shows the population distribution in the Scandinavian region. No known reproductions of this map exist.

3.4.2.2 John Snow's cholera map

Display of discrete data in the form of points in a map can provide convincing evidence related to medical geography. During the mid-1850s, cholera was a major concern. When a large outbreak occurred in London in 1854, Dr. John Snow created a dot distribution map that settled a debate between two schools of thought: that cholera is transmitted not through the inhalation of infected air, but through the ingestion of contaminated water or food.

Snow's map of the 1854 Broad Street cholera outbreak in London was simple and effective in its design. The base map is a simple road network, with few buildings named or depicted. The study area is outlined along the relevant road centerlines. Water pumps around the neighborhood are symbolized with points and bold, upper-case labels. Cholera deaths are depicted along the road network in their correct locations by address, with quantities measured by parallel tick marks stacked next to the road. The symbology, while simple, is effective in a study of fatal disease. The symbology of the cholera deaths is reminiscent of large Plague events, where bodies are stacked next to the roadway for disposal.

The map showed that a high number of deaths were occurring near a water pump on Broad Street at Cambridge Street. Snow petitioned the local authorities to remove the pump's handle, which caused dramatic decreases in cholera cases in the immediate area. The map helped the germ theory of disease transmission supplant miasma theory as the widely accepted view.

3.4.3 Advantages and disadvantages of dot distribution maps

Dot maps are advantageous when mapping phenomena that change smoothly over a space, as the dot maps will visually match the phenomena.

Dot distribution maps also have disadvantages. One such disadvantage is that the actual dot placement may be random. That is, there may be no actual phenomenon where the dots are located. Second, the

subjective nature of the dot size and spacing could give the map a biased view. Inappropriately sized or spaced dots can skew or distort the message a map attempts to communicate. If the dots are too numerous, it may be difficult for the reader to count the dots. This can cause the map to ineffective in communicating its message.

Solutions to problems in dot maps can be found through the combination of remotely sensed data. Satellite imagery showing lights at night can give good indication of population density. With such information, clustering of dots to relevant areas could possibly be applied to analyze. Data from the Oak Ridge National Laboratory's Landscan project is commonly used in this emerging technique.

3.4.4 Dot maps today

A large portion of the new maps appearing today is generated not through government agencies or geographical societies and associations, but by common individuals from all walks of life. Users of such virtual globe services as Google Earth constantly create new thematic maps quickly and easily. Such services offer a base platform upon which users can display layers as they choose, or even add their own data to the display.

One example was created on April 21, 2009 by Pittsburgh biochemist Harry Niman. The map shows where the 2009 H1N1 swine flu virus was spreading at the time. Despite the fact that the data shown was not backed by any official agency, the author's creation had "gone viral", surpassing 290,000 web views and 3,000 comments within nine days of its being published online.

Emerging technology is increasingly making mapping more mobile. Dot distribution maps, through their simple and effective displays, are becoming the standard for such on-the-fly mapping services.

3.5 **Flow-line maps**

Flow maps in map-making are by definition of Phan (2005) "a mix of maps and flow charts, that show the movement of objects from one location to another, such as the number of people in a migration, the amount of goods being traded, or the number of packets in a network".

3.5.1 Overview

Flow maps according to Harris (1999) "can be used to show movement of almost anything, including tangible things such as people, products, natural resources, weather, etc., as well as intangible things such as know-how, talent, credit of goodwill". Flow maps can indicate things like:

- What it is that flows, moves, migrates, etc.
- What direction the flow is moving and/or what the source and destination are.
- How much is flowing, being transferred, transported, etc.
- General information about what is flowing and how it is flowing.

In contrast to route maps, flow maps show little aside from the paths from one point to another.

3.5.2 Other types of flow maps

Beside the flow maps in map-making there are several other kinds of flow maps:

- Baker flow map of fluid flows
- Flow map or *solution operator*.
- Process flow map of a manufacturing process
- Sankey diagram in petroleum engineering
- Traffic Flow Maps
- XSL flow maps.

3.6 Map projection

A **map projection** is a systematic transformation of the latitudes and longitudes of locations on the surface of a sphere or an ellipsoid into locations on a plane. Map projections are necessary for creating maps. All map projections distort the surface in some fashion. Depending on the purpose of the map, some distortions are acceptable and others are not; therefore different map projections exist in order to preserve some properties of the sphere-like body at the expense of other properties. There is no limit to the number of possible map projections.

More generally, the surfaces of planetary bodies can be mapped even if they are too irregular to be modeled well with a sphere or ellipsoid. Even more generally, projections are the subject of several pure mathematical fields, including differential geometry and projective geometry. However "map projection" refers specifically to a cartographic projection.

3.6.1 Background

Maps can be more useful than globes in many situations: they are more compact and easier to store; they readily accommodate an enormous range of scales; they are viewed easily on computer displays; they can facilitate measuring properties of the terrain being mapped; they can show larger portions of the Earth's surface at once; and they are cheaper to produce and transport. These useful traits of maps motivate the development of map projections.

However, Carl Friedrich Gauss's Theorema Egregium proved that a sphere's surface cannot be represented on a plane without distortion. The same applies to other reference surfaces used as models for the Earth. Since any map projection is a representation of one of those surfaces on a plane, all map projections distort. Every distinct map projection distorts in a distinct way. The study of map projections is the characterization of these distortions.

Projection is not limited to perspective projections, such as those resulting from casting a shadow on a screen, or the rectilinear image produced by a pinhole camera on a flat film plate. Rather, any mathematical function transforming coordinates from the curved surface to the plane is a projection. Few projections in actual use are perspective.

For simplicity most of this article assumes that the surface to be mapped is that of a sphere. In reality, the Earth and other large celestial bodies are generally better modeled as oblate spheroids, whereas small objects such as asteroids often have irregular shapes. These other surfaces can be mapped as well. Therefore, more generally, a map projection is any method of "flattening" into a plane a continuous curved surface.

3.6.2 Metric properties of maps

Many properties can be measured on the Earth's surface independently of its geography. Some of these properties are:

- Area
- Shape
- Direction
- Bearing
- Distance
- Scale

Map projections can be constructed to preserve one or more of these properties, though not all of them simultaneously. Each projection preserves or compromises or approximates basic metric properties in different ways. The purpose of the map determines which projection should form the base for the map. Because many purposes exist for maps, many projections have been created to suit those purposes.

Another consideration in the configuration of a projection is its compatibility with data sets to be used on the map. Data sets are geographic information; their collection depends on the chosen datum (model) of the Earth. Different datums assign slightly different coordinates to the same location, so in large scale maps, such as those from national mapping systems, it is important to match the datum to the projection. The slight differences in coordinate assignation between different datums is not a concern for world maps or other vast territories, where such differences get shrunk to imperceptibility.

3.6.3 Which projection is best?

The mathematics of projection does not permit any particular map projection to be "best" for everything. Something will always get distorted. Therefore a diversity of projections exists to service the many uses of maps and their vast range of scales.

Modern national mapping systems typically employ a transverse Mercator or close variant for large-scale maps in order to preserve conformality and low variation in scale over small areas. For smaller-scale maps, such as those spanning continents or the entire world, many projections are in common use according to their fitness for the purpose.

Thematic maps normally require an equal area projection so that phenomena per unit area are shown in correct proportion. However, representing area ratios correctly necessarily distorts shapes more than many maps that are not equal-area. Hence reference maps of the world often appear on compromise projections instead. Due to the severe distortions inherent in any map of the world, within reason the choice of projection becomes largely one of æsthetics.

The Mercator projection, developed for navigational purposes, has often been used in world maps where other projections would have been more appropriate. This problem has long been recognized even outside professional circles. For example a 1943 New York Times editorial states:

The time has come to discard [the Mercator] for something that represents the continents and directions less deceptively... Although its usage... has diminished... it is still highly popular as a wall map apparently in part because, as a rectangular map, it fills a rectangular wall space with more map, and clearly because its familiarity breeds more popularity.

A controversy in the 1980s over the Peters map motivated the American Cartographic Association (now Map-making and Geographic Information Society) to produce a series of booklets (including *Which Map is Best*) designed to educate the public about map projections and distortion in maps. In 1989 and 1990, after some internal debate, seven North American geographic organizations adopted a resolution recommending against using any rectangular projection (including Mercator and Gall–Peters) for reference maps of the world.

3.6.4 Construction of a map projection

The creation of a map projection involves two steps:

1. Selection of a model for the shape of the Earth or planetary body (usually choosing between a sphere or ellipsoid). Because the Earth's actual shape is irregular, information is lost in this step.
2. Transformation of geographic coordinates (longitude and latitude) to Cartesian (x, y) or polar plane coordinates. Cartesian coordinates normally have a simple relation to eastings and northings defined on a grid superimposed on the projection.

Some of the simplest map projections are literally projections, as obtained by placing a light source at some definite point relative to the globe and projecting its features onto a specified surface. This is not the case for most projections, which are defined **only** in terms of mathematical formulae that have no direct geometric interpretation.

3.6.5 Choosing a projection surface

A surface that can be unfolded or unrolled into a plane or sheet without stretching, tearing or shrinking is called a *developable surface*. The cylinder, cone and of course the plane are all developable surfaces. The sphere and ellipsoid do not have developable surfaces, so any projection of them onto a plane will have to distort the image. (To compare, one cannot flatten an orange peel without tearing and warping it.)

One way of describing a projection is first to project from the Earth's surface to a developable surface such as a cylinder or cone, and then to unroll the surface into a plane. While the first step inevitably distorts some properties of the globe, the developable surface can then be unfolded without further distortion.

3.6.6 Aspects of the projection

Once a choice is made between projecting onto a cylinder, cone, or plane, the **aspect** of the shape must be specified. The aspect describes how the developable surface is placed relative to the globe: it may be *normal* (such that the surface's axis of symmetry coincide with the Earth's axis), *transverse* (at right angles to the Earth's axis) or *oblique* (any angle in between). The developable surface may also be either *tangent* or *secant* to the sphere or ellipsoid. Tangent means the surface touches but does not slice through the globe; secant means the surface does slice through the globe. Moving the developable surface away from contact with the globe never preserves or optimizes metric properties, so that possibility is not discussed further here.

3.6.7 Scale

A globe is the only way to represent the earth with constant scale throughout the entire map in all directions. A map cannot achieve that property in any area, no matter how small. It can, however, achieve constant scale along specific lines.

Some possible properties are:

- The scale depends on location, but not on direction. This is equivalent to preservation of angles, the defining characteristic of a conformal map.
- Scale is constant along any parallel in the direction of the parallel. This applies for any cylindrical or pseudocylindrical projection in normal aspect.
- The combination of the above: the scale depends on latitude only, not on longitude or direction. This applies for the Mercator projection in normal aspect.
- Scale is constant along all straight lines radiating from a particular geographic location. This is the defining characteristic of an equidistant projection such as the Azimuthal equidistant projection. There are also projections (Maurer, Close) where true distances from *two* points are preserved.

3.6.8 Choosing a model for the shape of the Earth

Projection construction is also affected by how the shape of the Earth is approximated. However, the Earth's actual shape is closer to an oblate ellipsoid. Whether spherical or ellipsoidal, the principles discussed hold without loss of generality.

Selecting a model for a shape of the Earth involves choosing between the advantages and disadvantages of a sphere versus an ellipsoid. Spherical models are useful for small-scale maps such as world atlases and globes, since the error at that scale is not usually noticeable or important enough to justify using the more complicated ellipsoid. The ellipsoidal model is commonly used to construct topographic maps and for other large- and medium-scale maps that need to accurately depict the land surface.

A third model of the shape of the Earth is the geoid, a complex and more accurate representation of the global mean sea level surface that is obtained through a combination of terrestrial and satellite gravity measurements. This model is not used for mapping because of its complexity, but rather is used for control purposes in the construction of geographic datums. (In geodesy, plural of "datum" is "datums" rather than "data".) A geoid is used to construct a datum by adding irregularities to the ellipsoid in order to better match the Earth's actual shape. It takes into account the large-scale features in the Earth's gravity field associated with mantle convection patterns, and the gravity signatures of very large geomorphic features such as mountain ranges, plateaus and plains.

Historically, datums have been based on ellipsoids that best represent the geoid within the region that the datum is intended to map. Controls (modifications) are added to the ellipsoid in order to construct the datum, which is specialized for a specific geographic region (such as the North American Datum). A few modern datums, such as WGS84 which is used in the Global Positioning System, are optimized to represent the entire earth as well as possible with a single ellipsoid, at the expense of accuracy in smaller regions.

3.6.9 Classification

A fundamental projection classification is based on the type of projection surface onto which the globe is conceptually projected. The projections are described in terms of placing a gigantic surface in contact with the earth, followed by an implied scaling operation. These surfaces are cylindrical (e.g. Mercator), conic (e.g., Albers), or azimuthal or plane (e.g. stereographic). Many mathematical projections, however, do not neatly fit into any of these three conceptual projection methods. Hence other peer categories have been described in the literature, such as pseudoconic, pseudocylindrical, pseudoazimuthal, retroazimuthal, and polyconic.

Another way to classify projections is according to properties of the model they preserve. Some of the more common categories are:

- Preserving direction (*azimuthal*), a trait possible only from one or two points to every other point
- Preserving shape locally (*conformal* or *orthomorphic*)
- Preserving area (*equal-area* or *equiareal* or *equivalent* or *authalic*)
- Preserving distance (*equidistant*), a trait possible only between one or two points and every other point
- Preserving shortest route, a trait preserved only by the gnomonic projection

Because the sphere is not a developable surface, it is impossible to construct a map projection that is both equal-area and conformal.

3.6.10 Projections by surface

The three developable surfaces (plane, cylinder, cone) provide useful models for understanding, describing, and developing map projections. However, these models are limited in two fundamental ways. For one thing, most world projections in actual use do not fall into any of those categories. For another thing, even most projections that do fall into those categories are not naturally attainable through physical projection.

No reference has been made in the above definitions to cylinders, cones or planes. The projections are termed cylindrical or conic because they can be regarded as developed on a cylinder or a cone, as the case may be, but it is as well to dispense with picturing cylinders and cones, since they have given rise to much misunderstanding. Particularly is this so with regard to the conic projections with two standard parallels: they may be regarded as developed on cones, but they are cones which bear no simple relationship to the sphere. In reality, cylinders and cones provide us with convenient descriptive terms, but little else.

Lee's objection refers to the way the terms *cylindrical*, *conic*, and *planar* (azimuthal) have been abstracted in the field of map projections. If maps were projected as in light shining through a globe onto a developable surface, then the spacing of parallels would follow a very limited set of possibilities. Such a cylindrical projection (for example) is one which:

1. Is rectangular;
2. Has straight vertical meridians, spaced evenly;
3. Has straight parallels symmetrically placed about the equator;
4. Has parallels constrained to where they fall when light shines through the globe onto the cylinder, with the light source someplace along the line formed by the intersection of the prime meridian with the equator, and the center of the sphere.

(If you rotate the globe before projecting then the parallels and meridians will not necessarily still be straight lines. Rotations are normally ignored for the purpose of classification.)

Where the light source emanates along the line described in this last constraint is what yields the differences between the various "natural" cylindrical projections. But the term *cylindrical* as used in the field of map projections relaxes the last constraint entirely. Instead the parallels can be placed according to any algorithm the designer has decided suits the needs of the map. The famous Mercator projection is one in which the placement of parallels does not arise by "projection"; instead parallels are placed how they need to be in order to satisfy the property that a course of constant bearing is always plotted as a straight line.

3.6.11 Cylindrical

The term "normal cylindrical projection" is used to refer to any projection in which meridians are mapped to equally spaced vertical lines and circles of latitude (parallels) are mapped to horizontal lines.

The mapping of meridians to vertical lines can be visualized by imagining a cylinder whose axis coincides with the Earth's axis of rotation. This cylinder is wrapped around the Earth, projected onto, and then unrolled.

By the geometry of their construction, cylindrical projection stretch distances east-west. The amount of stretch is the same at any chosen latitude on all cylindrical projections, and is given by the secant of the latitude as a multiple of the equator's scale. The various cylindrical projections are distinguished from each other solely by their north-south stretching (where latitude is given by φ):

- North-south stretching equals east-west stretching : The east-west scale matches the north-south scale: conformal cylindrical or Mercator; this distorts areas excessively in high latitudes.
- North-south stretching grows with latitude faster than east-west stretching: The cylindrical perspective (or central cylindrical) projection; unsuitable because distortion is even worse than in the Mercator projection.
- North-south stretching grows with latitude, but less quickly than the east-west stretching: such as the Miller cylindrical projection.
- North-south distances neither stretched nor compressed (1): equirectangular projection or "plate carrée".
- North-south compression precisely the reciprocal of east-west stretching: equal-area cylindrical. This projection has many named specializations differing only in the scaling constant. Some of those specializations are the Gall–Peters or Gall orthographic, Behrmann, and Lambert cylindrical equal-area). This kind of projection divides north-south distances by a factor equal to the secant of the latitude, preserving area at the expense of shapes.

In the first case (Mercator), the east-west scale always equals the north-south scale. In the second case (central cylindrical), the north-south scale exceeds the east-west scale everywhere away from the equator. Each remaining case has a pair of secant lines—a pair of identical latitudes of opposite sign (or else the equator) at which the east-west scale matches the north-south-scale.

Normal cylindrical projection map the whole Earth as a finite rectangular, except in the first two cases, where the rectangle stretches infinitely tall while retaining constant width.

3.6.12 Pseudocylindrical

Pseudocylindrical projections represent the *central* meridian as a straight line segment. Other meridians are longer than the central meridian and bow outward away from the central meridian. Pseudocylindrical projections map parallels as straight lines. Along parallels, each point from the surface is mapped at a distance from the central meridian that is proportional to its difference in longitude from the central meridian. On a pseudocylindrical map, any point further from the equator than some other point has a higher latitude than the other point, preserving north-south relationships. This trait is useful when illustrating phenomena that depend on latitude, such as climate. Examples of pseudocylindrical projections include:

- Sinusoidal, which was the first pseudocylindrical projection developed. Vertical scale and horizontal scale are the same throughout, resulting in an equal-area map. On the map, as in reality, the length of each parallel is proportional to the cosine of the latitude. Thus the shape of the map of the whole earth is the region between two symmetric rotated cosine curves. The true distance between two points on the same meridian corresponds to the distance on the map between the two parallels, which is smaller than the distance between the two points on the map. The distance between two points on the same parallel is true. The area of any region is true.
- Collignon projection, which in its most common forms represents each meridian as 2 straight line segments, one from each pole to the equator.

3.6.13 Hybrid

The HEALPix projection combines an equal-area cylindrical projection in equatorial regions with the Collignon projection in polar areas.

3.6.14 Conic

The term "conic projection" is used to refer to any projection in which meridians are mapped to equally spaced lines radiating out from the apex and circles of latitude (parallels) are mapped to circular arcs centered on the apex.

When making a conic map, the map maker arbitrarily picks two standard parallels. Those standard parallels may be visualized as secant lines where the cone intersects the globe—or, if the map maker chooses the same parallel twice, as the tangent line where the cone is tangent to the globe. The resulting conic map has low distortion in scale, shape, and area near those standard parallels. Distances along the parallels to the north of both standard parallels or to the south of both standard parallels are stretched; distances along parallels between the standard parallels are compressed. When a single standard parallel is used, distances along all other parallels are stretched.

The most popular conic maps include:

- Equidistant conic, which keeps parallel evenly spaced along the meridians to preserve a constant distance scale along each meridian, typically the same or similar scale as along the standard parallels.
- Albers conic, which adjusts the north-south distance between non-standard parallels to compensate for the east-west stretching or compression, giving an equal-area map.
- Lambert conformal conic, which adjusts the north-south distance between non-standard parallels to equal the east-west stretching, giving a conformal map.

3.6.15 Pseudoconic

- Bonn
- Werner cordiform, upon which distances are correct from one pole, as well as along all parolees.
- Continuous American polyconic

3.6.16 Azimuthal (projections onto a plane)

Azimuthal projections have the property that directions from a central point are preserved and therefore great circles through the central point are represented by straight lines on the map. Usually these projections also have radial symmetry in the scales and hence in the distortions: map distances from the central point are computed by a function r (d) of the true distanced, independent of the angle; correspondingly, circles with the central point as a center are mapped into circles which have as center the central point on the map.

The mapping of radial lines can be visualized by imagining a plane tangent to the Earth, with the central point as tangent point.

The radial scale is'(d) and the transverse scale *or* (d) /(R sin (d/R)) where R is the radius of the Earth.

Some azimuthal projections are true perspective projections; that is, they can be constructed mechanically, projecting the surface of the Earth by extending lines from a point of perspective (along an infinite line through the tangent point and the tangent point's antipode) onto the plane:

- The gnomonic projection displays great circles as straight lines. Can be constructed by using a point of perspective at the center of the Earth. $r(d) = c \tan(d/R)$; a hemisphere already requires an infinite map,
- The General Perspective projection can be constructed by using a point of perspective outside the earth. Photographs of Earth (such as those from the International Space Station) give this perspective.
- The orthographic projection maps each point on the earth to the closest point on the plane. Can be constructed from a point of perspective an infinite distance from the tangent point; $r(d) = c \sin(d/R)$. Can display up to a hemisphere on a finite circle. Photographs of Earth from far enough away, such as the Moon, give this perspective.
- The azimuthal conformal projection, also known as the stereographic projection, can be constructed by using the tangent point's antipode as the point of perspective. $r(d) = c \tan(d/2R)$; the scale is $c/(2R \cos^2(d/2R))$. Can display nearly the entire sphere's surface on a finite circle. The sphere's full surface requires an infinite map.

Other azimuthal projections are not true perspective projections:

- Azimuthal equidistant: $r(d) = cd$; it is used by amateur radio operators to know the direction to point their antennas toward a point and see the distance to it. Distance from the tangent point on the map is proportional to surface distance on the earth (; for the case where the tangent point is the North Pole.
- Lambert azimuthal equal-area. Distance from the tangent point on the map is proportional to straight-line distance through the earth: $r(d) = c \sin(d/2R)$
- Logarithmic azimuthal is constructed so that each point's distance from the center of the map is the logarithm of its distance from the tangent point on the Earth. $r(d) = c \ln(d/d_0)$; locations closer than at a distance equal to the constant d_0 are not shown.

3.6.17 Projections by preservation of a metric property

3.6.17.1 Conformal

Conformal, or orthomorphic, map projections preserve angles locally, implying that they map infinitesimal circles of constant size anywhere on the Earth to infinitesimal circles of varying sizes on the map. In contrast, mappings that are not conformal distort most such small circles into ellipses of distortion. An important consequence of conformality is that relative angles at each point of the map are correct, and the local scale (although varying throughout the map) in every direction around any one point is constant. These are some conformal projections:

- Mercator: Rhumb lines are represented by straight segments
- Transverse Mercator
- Stereographic: Any circle of a sphere, great and small, maps to a circle or straight line.
- Roussilhe
- Lambert conformal conic
- Peirce quincuncial projection
- Adams hemisphere-in-a-square projection
- Guyou hemisphere-in-a-square projection

3.6.17.2 Equal-area

These are some projections that preserve area:

- Gall orthographic (also known as Gall–Peters, or Peters, projection)
- Albers conic
- Lambert azimuthal equal-area
- Lambert cylindrical equal-area
- Mollweide
- Hammer
- Briesemeister
- Sinusoidal
- Werner
- Bonne
- Bottomley
- Goode's homolosine
- Hobo–Dyer
- Collignon
- Tobler hyperelliptical
- Snyder’s equal-area polyhedral projection, used for geodesic grids.

3.6.17.3 Equidistant

These are some projections that preserve distance from some standard point or line:

- Equirectangular—distances along meridians are conserved
- Plate carrée—an Equirectangular projection centered at the equator
- Azimuthal equidistant—distances along great circles radiating from centre are conserved
- Equidistant conic
- Sinusoidal—distances along parallels are conserved

- Werner cordiform distances from the North Pole are correct as are the curved distance on parallels
- Soldner
- Two-point equidistant: two "control points" are arbitrarily chosen by the map maker. Distance from any point on the map to each control point is proportional to surface distance on the earth.

3.6.17.4 Gnomonic

Great circles are displayed as straight lines:

- Gnomonic projection

3.6.17.5 Retroazimuthal

Direction to a fixed location B (the bearing at the starting location A of the shortest route) corresponds to the direction on the map from A to B:

- Littrow—the only conformal retroazimuthal projection
- Hammer retroazimuthal—also preserves distance from the central point
- Craig retroazimuthal *aka* Mecca or Qibla—also has vertical meridians

3.6.17.6 Compromise projections

Compromise projections give up the idea of perfectly preserving metric properties, seeking instead to strike a balance between distortions, or to simply make things "look right". Most of these types of projections distort shape in the polar regions more than at the equator. These are some compromise projections:

- Robinson
- van der Grinten
- Miller cylindrical
- Winkel Tripel
- Buckminster Fuller's Dymaxion
- B.J.S. Cahill's Butterfly Map
- Kavrayskiy VII
- Wagner VI projection
- Chamberlin trimetric
- Oronce Finé's cordiform

3.7 **Digital mapping**

Digital mapping (also called digital map-making) is the process by which a collection of data is compiled and formatted into a virtual image. The primary function of this technology is to produce maps that give accurate representations of a particular area, detailing major road arteries and other points of interest. The technology also allows the calculation of distances from once place to another.

Though digital mapping can be found in a variety of computer applications, such as Google Earth, the main use of these maps is with the Global Positioning System, or GPS satellite network, used in standard automotive navigation systems.

3.7.1 History

3.7.1.1 From Paper to Paperless

The roots of digital mapping lie within traditional paper maps such as the Thomas Guide. Paper maps provide basic landscapes similar to digitized road maps, yet are often cumbersome, cover only a designated area, and lack many specific details such as road blocks. In addition, there is no way to "update" a paper map except to obtain a new version. On the other hand, digital maps, in many cases, can be updated through synchronization with updates from company servers.

3.7.1.2 Expanded Capabilities

Early digital maps had the same basic functionality as paper maps—that is, they provided a "virtual view" of roads generally outlined by the terrain encompassing the surrounding area. However, as digital maps have grown with the expansion of GPS technology in the past decade, live traffic updates, points of interest and service locations have been added to enhance digital maps to be more "user conscious." Traditional "virtual views" are now only part of digital mapping. In many cases, users can choose between virtual maps, satellite (aerial views), and hybrid (a combination of virtual map and aerial views) views. With the ability to update and expand digital mapping devices, newly constructed roads and places can be added to appear on maps.

3.7.1.3 Data Collection

Digital maps heavily rely upon a vast amount of data collected over time. Most of the information that comprise digital maps is the culmination of satellite imagery as well as street level information. Maps must be updated frequently to provide users with the most accurate reflection of a location. While there is a wide spectrum of companies that specialize in digital mapping, the basic premise is that digital maps will accurately portray roads as they actually appear to give "lifelike experiences."

3.7.2 Functionality and Use

3.7.2.1 Computer Applications

Computer programs and applications such as Google Earth and Google Maps provide map views from space and street level of much of the world. Used primarily for recreational use, Google Earth provides digital mapping in personal applications, such as tracking distances or finding locations.

3.7.2.2 Scientific Applications

The development of mobile computing (PDAs, tablet PCs, laptops, etc.) has recently (since about 2000) spurred the use of digital mapping in the sciences and applied sciences. As of 2009, science fields that use digital mapping technology include geology, engineering, architecture, land surveying, mining, forestry, environmental, and archaeology.

3.7.3 GPS Navigation Systems

The principle use by which digital mapping has grown in the past decade has been its connection to Global Positioning System (GPS) technology. GPS is the foundation behind digital mapping navigation systems.

3.7.4 How It Works

The coordinates and position as well as atomic time obtained by a terrestrial GPS receiver from GPS satellites orbiting Earth interact together to provide the digital mapping programming with points of origin in addition to the destination points needed to calculate the distance. This information is then analyzed and compiled to create a map that provides the easiest and most efficient way to reach a destination.

More technically speaking, the device operates in the following manner:

1. GPS receivers collect data from at least four GPS satellites orbiting the Earth, calculating position in three dimensions.
2. The GPS receiver then utilizes position to provide GPS coordinates, or exact points of latitudinal and longitudinal direction from GPS satellites.
3. The points, or coordinates, output an accurate range between approximately "10-20 meters" of the actual location.
4. The beginning point, entered via GPS coordinates, and the ending point, (address or coordinates) input by the user, is then entered into the digital map.
5. The map outputs a real-time visual representation of the route. The map then moves along the path of the driver.
6. If the driver drifts from the designated route, the navigation system will use the current coordinates to recalculate a route to the destination location.

3.8 Ebstorf Map

The **Ebstorf Map** is an example of a mappa mundi (a Medieval European map of the world) similar to the Hereford Map. It was made by Gervase of Ebstorf, who was possibly the same man as Gervase of Tilbury, some time in the thirteenth century.

The map was found in a convent in Ebstorf, in northern Germany, in 1843. It was a very large map, painted on 30 goatskins sewn together and measuring around 3.6 by 3.6 meters (12 ft × 12 ft) —a greatly elaborated version of the common medieval tripartite, or T and O, map, centered on Jerusalem with east at top. The head of Christ was depicted at the top of the map, with his hands on either side and his feet at the bottom. Rome is represented in the shape of a lion, and the map reflects an evident interest in the distribution of bishoprics.

There was text around the map, which included descriptions of animals, the creation of the world, definitions of terms, and a sketch of the more common sort of T and O map with an explanation of how the world is divided into three parts. The map incorporated both pagan and biblical history.

The arguments for Gervase of Tilbury's being the mapmaker are based on the name Gervase, which was an uncommon name in Northern Germany at the time, and on some similarities between the world views

of the mapmaker and Gervase of Tilbury. The editors of the Oxford Medieval Texts edition of Gervase of Tilbury's *Otia Imperialia* conclude that despite there being the same man is an "attractive possibility", to accept it requires "too many improbable assumptions".

The original was destroyed in 1943, during the bombing of Hanover in World War II. There survives a set of black-and-white photographs of the original map, taken in 1891, and several color facsimiles of it was made before it was destroyed.

3.9 Scale (map)

The **scale** of a map is the ratio of a distance on the map to the corresponding distance on the ground. This simple concept is complicated by the curvature of the Earth's surface, which forces scale to vary across a map. Because of this variation, the concept of scale becomes meaningful in two distinct ways. The first way is the ratio of the size of the **generating globe** to the size of the Earth. The generating globe is a conceptual model to which the Earth is shrunk and from which the map is projected.

The ratio of the Earth's size to the generating globe's size is called the **nominal scale** (= **principal scale** = **representative fraction**). Many maps state the nominal scale and may even display a bar scale (sometimes merely called a 'scale') to represent it. The second distinct concept of scale applies to the variation in scale across a map. It is the ratio of the mapped point's scale to the nominal scale. In this case 'scale' means the **scale factor** (= **point scale** = **particular scale**).

If the region of the map is small enough to ignore Earth's curvature—a town plan, for example—then a single value can be used as the scale without causing measurement errors. In maps covering larger areas, or the whole Earth, the map's scale may be less useful or even useless in measuring distances. The map projection becomes critical in understanding how scale varies throughout the map. When scale varies noticeably, it can be accounted for as the scale factor. Tissot's indicatrix is often used to illustrate the variation of point scale across a map.

3.9.1 **The terminology of scales**

3.9.1.1 **Representation of scale**

Map scales may be expressed in words (a lexical scale), as a ratio, or as a fraction. Examples are:

> 'one centimetre to one hundred metres' or 1:10,000 or 1/10,000
> 'one inch to one mile' or 1:63,360 or 1/63,360
> 'one centimetre to one thousand kilometres' or 1:100,000,000 or 1/100,000,000. (The ratio would usually be abbreviated to 1:100M

3.9.1.2 **Bar scale vs. lexical scale**

In addition to the above many maps carry one or more *(graphical)* **bar scales**. For example some British maps presently (2009) use three bar scales for kilometers, miles and nautical miles.

A lexical scale on a recently published map, in a language known to the user, may be easier for a non-mathematician to visualize than a ratio: if the scale is an inch to two miles and he can see that two villages are about two inches apart on the map then it is easy to work out that they are about four miles apart on the ground.

On the other hand, a lexical scale may cause problems if it expressed in a language that the user does not understand or in obsolete or ill-defined units. On the other hand ratios and fractions may be more acceptable to the numerate user since they are immediately accessible in any language. For example a scale of one inch to a furlong (1:7920) will be understood by many older people in countries where Imperial units used to be taught in schools. But a scale of one pouce to one league may be about 1:144,000 but it depends on the cartographer's choice of the many possible definitions for a league, and only a minority of modern users will be familiar with the units used.

3.9.1.3 **Large scale, medium scale, small scale**

Maps are often described as **small scale**, typically for world maps or large regional maps, showing large areas of land on a small space, or **large scale**, showing smaller areas in more detail, typically for county maps or town plans. The town plan might be on a scale of 1:10,000 and the world map might be on a scale of 1:100,000,000. The following table describes typical ranges for these scales but should not be considered authoritative because there is no standard:

Classification	**Range**	**Examples**
large scale	1:0 – 1:600,000	1:0.00001 for map of virus; 1:5,000 for walking map of town
medium scale	1:600,000 – 1:2,000,000	Map of a country
small scale	1:2,000,000 – 1:∞	1:50,000,000 for world map; 1:10 for map of galaxy

The terms are sometimes used in the absolute sense of the table, but other times in a relative sense. For example, a map reader whose work refers solely to large-scale maps (as tabulated above) might refer to a map at 1:500,000 as small-scale.

3.9.1.4 **Scale variation**

Mapping large areas cause noticeable distortions due to flattening the significantly curved surface of the earth. How distortion gets distributed depends on the map projection. Scale varies across the map, and the stated map scale will only be an approximation.

3.9.1.5 **Large-scale maps with curvature neglected**

The region over which the earth can be regarded as flat depends on the accuracy of the survey measurements. If measured only to the nearest meter, then the curvature of the earth is undetectable over a meridian distance of about 100 kilometers (62 mi) and over an east-west line of about 80 km (at a latitude of 45 degrees). If surveyed to the nearest 1 millimeter (0.039 in), then the curvature is undetectable over a meridian distance of about 10 km and over an east-west line of about 8 km. Thus a city plan of New York accurate to one meter or a building site plan accurate to one millimeter would both satisfy the above conditions for the neglect of curvature. They can be treated by plane surveying and mapped by scale drawings in which any two points at the same distance on the drawing are at the same distance on the ground. True ground distances are calculated by measuring the distance on the map and then multiplying by the inverse of the scale fraction or, equivalently, simply using dividers to transfer the separation between the points on the map to a bar scale on the map.

3.9.1.6 **Point scale (or particular scale)**

As proved by Gauss's *Theorema Egregium*, a sphere (or ellipsoid) cannot be projected onto the plane without distortion. This is commonly illustrated by the impossibility of smoothing an orange peel onto a

flat surface without tearing and deforming it. The only true representation of a sphere at constant scale is another sphere such as the schoolroom globe.

Given the limited practical size of globes, we must use maps for detailed mapping. Maps require projections. A projection implies distortion: A constant separation on the map does not correspond to a constant separation on the ground. While a map may display a graphical bar scale, the scale must be used with the understanding that it will be accurate only on some lines of the map.

Let P be a point at latitude $\phi$and longitude λon the sphere (or ellipsoid). Let Q be a neighbouring point and let αbe the angle between the element PQ and the meridian at P: this angle is the **azimuth** angle of the element PQ. Let P' and Q' be corresponding points on the projection. The angle between the direction P'Q' and the projection of the meridian is the **bearing** β. In general $\alpha \neq \beta$. Comment: this precise distinction between azimuth (on the Earth's surface) and bearing (on the map) is not universally observed, many writers using the terms almost interchangeably.

Definition: the **point scale** at P is the ratio of the two distances P'Q' and PQ in the limit that Q approaches P. We write this as

$$\mu(\lambda,\, \phi,\, \alpha) = \lim_{Q \to P} \frac{P'Q'}{PQ},$$

where the notation indicates that the point scale is a function of the position of P and also the direction of the element PQ.

Definition: if P and Q lie on the same meridian $(\alpha = 0)$, the **meridian scale** is denoted by $h(\lambda,\, \phi)$.

Definition: if P and Q lie on the same parallel $(\alpha = \pi/2)$, the **parallel scale** is denoted by $k(\lambda,\, \phi)$.

Definition: if the point scale depends only on position, not on direction, we say that it is isotropic and conventionally denote its value in any direction by the parallel scale factor $k(\lambda, \phi)$.

Definition: A map projection is said to be conformal if the angle between a pair of lines intersecting at a point P is the same as the angle between the projected lines at the projected point P', for all pairs of lines intersecting at point P. A conformal map has an isotropic scale factor. Conversely isotropic scale factors across the map imply a conformal projection.

The isotropy of scale implies that *small* elements are stretched equally in all directions, that is the shape of a small element is preserved. This is the property of **orthomorphism** (from Greek 'right shape'). The qualification 'small' means that at any given accuracy of measurement no change can be detected in the scale factor over the element. Since conformal projections have an isotropic scale factor they have also been called **orthomorphic projections**. For example the Mercator projection is conformal since it is constructed to preserve angles and its scale factor is isotropic, a function of latitude only: Mercator *does* preserve shape in small regions.

Definition: on a conformal projection with an isotropic scale, points which have the same scale value may be joined to form the **isoscale lines**. These are not plotted on maps for end users but they feature in many of the standard texts.

3.9.1.7 **The representative fraction (RF) or principal scale**

There are two conventions used in setting down the equations of any given projection. For example, the equirectangular cylindrical projection may be written as

cartographers: $x = a\lambda \quad x = a\phi$
mathematicians: $x = \lambda \quad x = \phi$

Here we shall adopt the first of these conventions (following the usage in the surveys by Snyder). Clearly the above projection equations define positions of a huge cylinder wrapped around the Earth and then unrolled. We say that these coordinates define the **projection map** which must be distinguished logically from the actual **printed** (or viewed) maps. If the definition of point scale in the previous section is in terms of the projection map then we can expect the scale factors to be close to unity. For normal tangent cylindrical projections the scale along the equator is k=1 and in general the scale changes as we move off the equator. Analysis of scale on the projection map is an investigation of the change of k away from its true value of unity.

Actual **printed maps** are produced from the projection map by a *constant* scaling denoted by a ratio such as 1:100M (for whole world maps) or 1:10000 (for such as town plans). To avoid confusion in the use of the word 'scale' this constant scale fraction is called the **representative fraction** (RF) of the printed map and it is to be identified with the ratio printed on the map. The actual printed map coordinates for the equirectangular cylindrical projection are

printed map: $x = (RF)a\lambda \quad y = (RF)a\phi$

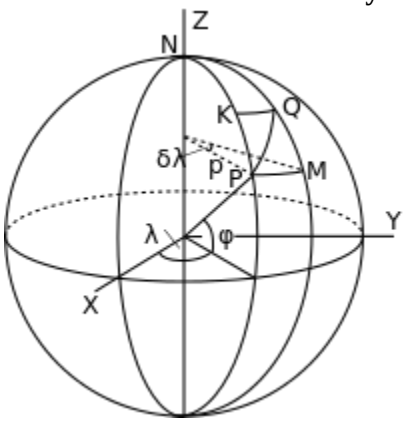
This convention allows a clear distinction of the intrinsic projection scaling and the reduction scaling.

From this point we ignore the RF and work with the projection map.

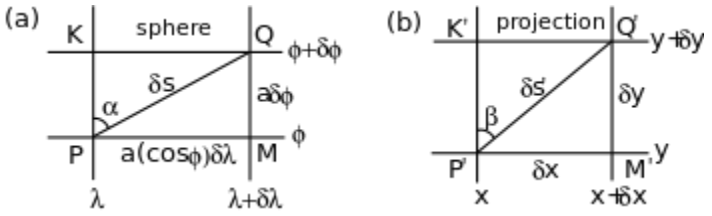
3.9.2 **Visualization of point scale: the Tissot indicatrix**

Consider a small circle on the surface of the Earth centered at a point P at latitude ϕ and longitude λ. Since the point scale varies with position and direction the projection of the circle on the projection will be distorted. Tissot proved that, as long as the distortion is not too great, the circle will become an ellipse on the projection. In general the dimension, shape and orientation of the ellipse will change over the projection. Superimposing these distortion ellipses on the map projection conveys the way in which the point scale is changing over the map. The distortion ellipse is known as Tissot's indicatrix. The example shown here is the Winkel tripel projection, the standard projection for world maps made by the National Geographic Society. The minimum distortion is on the central meridian at latitudes of 30 degrees (North and South).

3.9.3 **Point scale for normal cylindrical projections of the sphere**



The key to a *quantitative* understanding of scale is to consider an infinitesimal element on the sphere. The figure shows a point P at latitude ϕ and longitude λ on the sphere. The point Q is at latitude $\phi + \delta\phi$ and longitude $\lambda + \delta\lambda$. The lines PK and MQ are arcs of meridians of length $a\delta\phi$ where a is the radius of the sphere and ϕ is in radian measure. The lines PM and KQ are arcs of parallel circles of length $(a\cos\phi)\delta\lambda$ with λ in radian measure. In deriving a *point* property of the projection *at* P it suffices to take an infinitesimal element PMQK of the surface: in the limit of Q approaching P such an element tends to an infinitesimally small planar rectangle.



Infinitesimal elements on the sphere and a normal cylindrical projection

Normal cylindrical projections of the sphere have $x = a\lambda$ and y a function of latitude only. Therefore the infinitesimal element PMQK on the sphere projects to an infinitesimal element P'M'Q'K' which is an *exact* rectangle with a base $\delta x = a\delta\lambda$ and height δy. By comparing the elements on sphere and projection we can immediately deduce expressions for the scale factors on parallels and meridians. (We defer the treatment of the scale in a general direction to a mathematical addendum to this page.)

parallel scale factor $$k = \frac{\delta x}{a\cos\phi\,\delta\lambda} = \sec\phi$$

meridian scale factor $$h = \frac{\delta y}{a\delta\phi} = \frac{y'(\phi)}{a}$$

Note that the parallel scale factor $k = \sec\phi$ is independent of the definition of $y(\phi)$ so it is the same for all normal cylindrical projections. It is useful to note that

at latitude 30 degrees the parallel scale is $k = \sec 30^\circ = 2/\sqrt{3} = 1.15$
at latitude 45 degrees the parallel scale is $k = \sec 45^\circ = \sqrt{2} = 1.414$
at latitude 60 degrees the parallel scale is $k = \sec 60^\circ = 2$
at latitude 80 degrees the parallel scale is $k = \sec 80^\circ = 5.76$
at latitude 85 degrees the parallel scale is $k = \sec 85^\circ = 11.5$

The following examples illustrate three normal cylindrical projections and in each case the variation of scale with position and direction is illustrated by the use of Tissot's indicatrix.

3.9.4 **Three examples of normal cylindrical projection**

3.9.4.1**The equirectangular projection**

The equirectangular projection, also known as the Plate Carrée (French for "flat square") or (somewhat misleadingly) the equidistant projection, is defined by

$$x = a\lambda, \;\; y = a\phi,$$

where $a$is the radius of the sphere, $\lambda$is the longitude from the central meridian of the projection (here taken as the Greenwich meridian at $\lambda = 0$) and $\phi$is the latitude. Note that $\lambda$and $\phi$are in radians (obtained by multiplying the degree measure by a factor of π/180). The longitude $\lambda$is in the range $[-\pi, \pi]$and the latitude $\phi$is in the range $[-\pi/2, \pi/2]$.

Since $y'(\phi) = 1$the previous section gives

parallel scale, $k = \dfrac{\delta x}{a\cos\phi\,\delta\lambda} = \sec\phi$ meridian scale $h = \dfrac{\delta y}{a\delta\phi} = 1$

The figure illustrates the Tissot indicatrix for this projection. On the equator h=k=1 and the circular elements are undistorted on projection. At higher latitudes the circles are distorted into an ellipse given by stretching in the parallel direction only: there is no distortion in the meridian direction. The ratio of the major axis to the minor axis is $\sec\phi$. Clearly the area of the ellipse increases by the same factor.

It is instructive to consider the use of bar scales that might appear on a printed version of this projection. The scale is true (k=1) on the equator so that multiplying its length on a printed map by the inverse of the RF (or principal scale) gives the actual circumference of the Earth. The bar scale on the map is also drawn at the true scale so that transferring a separation between two points on the equator to the bar scale will give the correct distance between those points. The same is true on the meridians. On a parallel other than the equator the scale is $\sec\phi$so when we transfer a separation from a parallel to the bar scale we must divide the bar scale distance by this factor to obtain the distance between the points when measured along the parallel (which is not the true distance along a great circle). On a line at a bearing of say 45 degrees ($\beta = 45^\circ$) the scale is continuously varying with latitude and transferring a separation along the line to the bar scale does not give a distance related to the true distance in any simple way. Even if we could work out a distance along this line of constant bearing its relevance is questionable since such a line on

the projection corresponds to a complicated curve on the sphere. For these reasons bar scales on small-scale maps must be used with extreme caution.

3.9.4.2 **Mercator projection**

The Mercator projection maps the sphere to a rectangle (of infinite extent in the y-direction) by the equations

$$x = a\lambda$$
$$y = a \ln \left[\tan \left(\frac{\pi}{4} + \frac{\phi}{2} \right) \right]$$

where a, $\lambda$and $\phi$are as in the previous example. Since $y'(\phi) = a \sec \phi$the scale factors are:

parallel scale

$$k = \frac{\delta x}{a \cos \phi \, \delta \lambda} = \sec \phi$$

meridian scale

$$h = \frac{\delta y}{a \delta \phi} = \sec \phi$$

In the mathematical addendum below we prove that the point scale in an arbitrary direction is also equal to $\sec \phi$so the scale is isotropic (same in all directions), its magnitude increasing with latitude as $\sec \phi$. In the Tissot diagram each infinitesimal circular element preserves its shape but is enlarged more and more as the latitude increases.

3.9.4.3 **Lambert's equal area projection**

Lambert's equal area projection maps the sphere to a finite rectangle by the equations

$$x = a\lambda \qquad\qquad y = a \sin \phi$$

where a, $\lambda$and $\phi$are as in the previous example. Since $y'(\phi) = \cos \phi$the scale factors are

parallel scale $$k = \frac{\delta x}{a \cos \phi \, \delta \lambda} = \sec \phi$$

meridian scale $$h = \frac{\delta y}{a \delta \phi} = \cos \phi$$

The calculation of the point scale in an arbitrary direction is given below.

The vertical and horizontal scales now compensate each other (hk=1) and in the Tissot diagram each infinitesimal circular element is distorted into an ellipse of the *same* area as the undistorted circles on the equator.

3.9.4.4 **Graphs of scale factors**

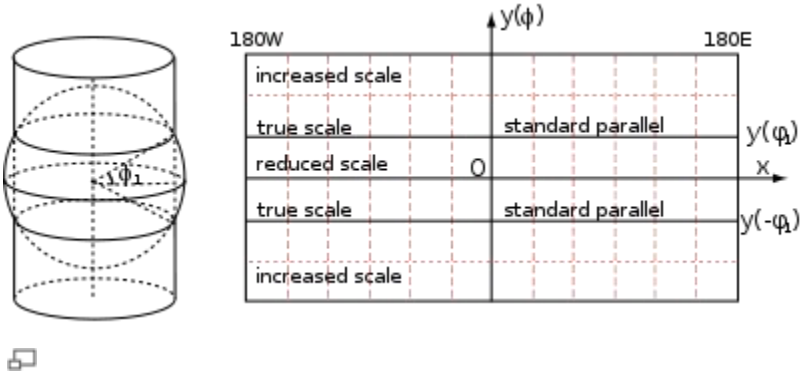
3.9.5 Scale variation on the Mercator projection

The Mercator point scale is unity on the equator because it is such that the auxiliary cylinder used in its construction is tangential to the Earth at the equator. For this reason the usual projection should be called a **tangent** projection. The scale varies with latitude as $k = \sec\phi$. Since $\sec\phi$tends to infinity as we approach the poles the Mercator map is grossly distorted at high latitudes and for this reason the projection is totally inappropriate for world maps (unless we are discussing navigation and rhumb lines). However, at a latitude of about 25 degrees the value of $\sec\phi$is about 1.1 so Mercator *is* accurate to within 10% in a strip of width 50 degrees centred on the equator. Narrower strips are better: a strip of width 16 degrees (centred on the equator) is accurate to within 1% or 1 part in 100.

A standard criterion for good large-scale maps is that the accuracy should be within 4 parts in 10,000, or 0.04%, corresponding to $k = 1.0004$. Since $\sec\phi$attains this value at $\phi = 1.62$degrees. Therefore the tangent Mercator projection is highly accurate within a strip of width 3.24 degrees centred on the equator. This corresponds to north-south distance of about 360 km (220 mi). Within this strip Mercator is *very* good, highly accurate and shape preserving because it is conformal (angle preserving). These observations prompted the development of the transverse Mercator projections in which a meridian is treated 'like an equator' of the projection so that we obtain an accurate map within a narrow distance of that meridian. Such maps are good for countries aligned nearly north-south (like Great Britain) and a set of 60 such maps is used for the Universal Transverse Mercator (UTM). Note that in both these projections (which are based on various ellipsoids) the transformation equations for x and y and the expression for the scale factor are complicated functions of both latitude and longitude.

3.9.6 Secant, or modified, projections

The basic idea of a secant projection is that the sphere is projected to a cylinder which intersects the sphere at two parallels, say $\phi_1$north and south. Clearly the scale is now true at these latitudes whereas parallels beneath these latitudes are contracted by the projection and their (parallel) scale factor must be less than one. The result is that deviation of the scale from unity is reduced over a wider range of latitudes.



As an example, one possible secant Mercator projection is defined by

$$x = 0.9996a\lambda \qquad\qquad y = 0.9996a\ln\left(\tan\left(\frac{\pi}{4}+\frac{\phi}{2}\right)\right).$$

The numeric multipliers do not alter the shape of the projection but it does mean that the scale factors are modified:

secant Mercator scale, $k = 0.9996\sec\phi.$

Thus

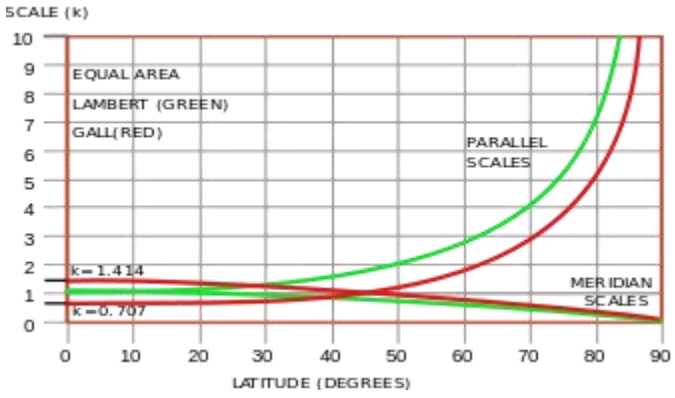
- the scale on the equator is 0.9996,
- the scale is k=1 at a latitude given by $\phi_1$where $\sec\phi_1 = 1/0.9996 = 1.00004$so that $\phi_1 = 1.62$degrees,
- k=1.0004 at a latitude $\phi_2$given by $\sec\phi_2 = 1.0004/0.9996 = 1.0008$for which $\phi_2 = 2.29$degrees. Therefore the projection has $1 < k < 1.0004$, that is an accuracy of 0.04%, over a wider strip of 4.58 degrees (compared with 3.24 degrees for the tangent form).

Such narrow zones of high accuracy are used in the UTM and the British OSGB projection, both of which are secant, transverse Mercator on the ellipsoid with the scale on the central meridian constant at $k_0 = 0.9996$. The isoscale lines with $k = 1$are slightly curved lines approximately 180 km east and west of the central meridian. The maximum value of the scale factor is 1.001 for UTM and 1.0007 for OSGB.

The lines of unit scale at latitude ϕ_1(north and south), where the cylindrical projection surface intersects the sphere, are the **standard parallels** of the secant projection.

Whilst a narrow band with $|k-1| < 0.0004$is important for high accuracy mapping at a large scale, for world maps much wider spaced standard parallels are used to control the scale variation. Examples are

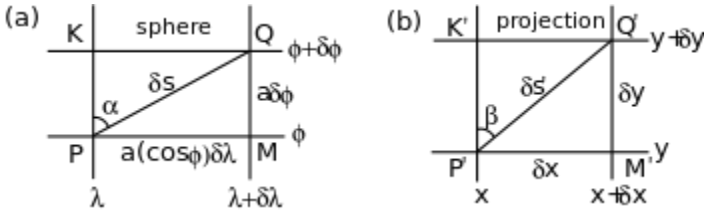
- Behrmann with standard parallels at 30N, 30S.
- Gall equal area with standard parallels at 45N, 45S.



Scale variation for the Lambert (green) and Gall (red) equal area projections.

The scale plots for the latter are shown below compared with the Lambert equal area scale factors. In the latter the equator is a single standard parallel and the parallel scale increases from k=1 to compensate the decrease in the meridian scale. For the Gall the parallel scale is reduced at the equator (to k=0.707) whilst the meridian scale is increased (to k=1.414). This gives rise to the gross distortion of shape in the Gall-Peters projection. (On the globe Africa is about as long as it is broad). Note that the meridian and parallel scales are both unity on the standard parallels.

3.9.7 Mathematical addendum



Infinitesimal elements on the sphere and a normal cylindrical projection

For normal cylindrical projections the geometry of the infinitesimal elements gives

$$\text{(a)} \quad \tan\alpha = \frac{a\cos\phi\,\delta\lambda}{a\,\delta\phi},$$

$$\text{(b)} \quad \tan\beta = \frac{\delta x}{\delta y} = \frac{a\delta\lambda}{\delta y}.$$

The relationship between the angles β and α is

$$\text{(c)} \quad \tan\beta = \frac{a\sec\phi}{y'(\phi)}\tan\alpha.$$

For the Mercator projection $y'(\phi) = a\sec\phi$ giving $\alpha = \beta$: angles are preserved. (Hardly surprising since this is the relation used to derive Mercator). For the equidistant and Lambert projections we have $y'(\phi) = a$ and $y'(\phi) = a\cos\phi$ respectively so the relationship between α and β depends upon the latitude ϕ. Denote the point scale at P when the infinitesimal element PQ makes an angle α with the meridian by μ_α. It is given by the ratio of distances:

$$\mu_\alpha = \lim_{Q\to P}\frac{P'Q'}{PQ} = \lim_{Q\to P}\frac{\sqrt{\delta x^2+\delta y^2}}{\sqrt{a^2\,\delta\phi^2 + a^2\cos^2\phi\,\delta\lambda^2}}.$$

Setting $\delta x = a\delta\lambda$ and substituting $\delta\phi$ and δy from equations (a) and (b) respectively gives

$$\mu_\alpha(\phi) = \sec\phi \left[\frac{\sin\alpha}{\sin\beta}\right].$$

For the projections other than Mercator we must first calculate β from α and ϕ using equation (c), before we can find μ_α. For example the equirectangular projection has $y' = a$ so that

$$\tan\beta = \sec\phi \tan\alpha.$$

If we consider a line of constant slope β on the projection both the corresponding value of α and the scale factor along the line are complicated functions of ϕ. There is no simple way of transferring a general finite separation to a bar scale and obtaining meaningful results.

3.10 Map coloring

Map coloring is the act of assigning different colors to different features on a map. There are two very different uses of this term. The first is in map-making, choosing the colors to be used when producing a map. The second is in mathematics, where the problem is to determine the minimum number of colors needed to color a map so that no two adjacent features have the same color.

3.10.1 Cartography

Color is a very useful attribute to depict different features on a map. Typical uses of color include displaying different political divisions, different elevations, or different kinds of roads. A choropleth map is a thematic map in which areas are colored differently to show the measurement of a statistical variable being displayed on the map. The choropleth map provides an easy way to visualize how a measurement varies across a geographic area or it shows the level of variability within a region.

Displaying the data in different hues can greatly affect the understanding or feel of the map. Also, the cartographer must take into account that many people have impaired color vision, and use colors that are easily distinguishable by these readers.

Colors can also be used to produce a three dimensional effect from two dimensional maps, either by explicit color-coding of the two images intended for different eyes, or by using the characteristics of the human visual system to make the map look three dimensional.

3.10.2 Mathematics

In mathematics there is a very strong link between map coloring and graph coloring, since every map showing different areas has a corresponding graph. By far the most famous result in this area is the four color theorem, which states that any planar map can be colored with at most four colors.

3.11 Map communication model

Map Communication Model is a theory in map-making that characterizes mapping as a process of transmitting geographic information via the map from the cartographer to the end-user.

3.11.1 Overview

By the mid-20th century, according to Crampton (2001) "cartographers as Arthur H. Robinson and others had begun to see the map as primarily a communication tool, and so developed a specific model for map communication, the map communication model (MCM)". This model, according to Andrews (1988) "can be grouped with the other major communication models of the time, such as the Shannon-Weaver and Lasswell models of communication. The map communication model led to a whole new body of research, methodologies and map design paradigms"

One of the implications of this communication model according to Donohue (2008) "endorsed an "epistemic break" that shifted our understandings of maps as communication systems to investigating them in terms of fields of power relations and exploring the "mapping environments in which knowledge is constructed"... This involved examining the social contexts in which maps were both produced and used, a departure from simply seeing maps as artifacts to be understood apart from this context". is a clear separation between cartographer and user, whereby the map was seen simply as an "intermediary between the cartographer and the user".

A second implication of this model is the presumption inherited from positivism that it is possible to separate facts from values. As Harley stated: Maps are never value-free images; except in the narrowest Euclidean sense they are not in themselves either true or false. Both in the selectivity of their content and in their signs and styles of representation maps are a way of conceiving, articulating, and structuring the human world which is biased towards, promoted by, and exerts influence upon particular sets of social relations. By accepting such premises it becomes easier to see how appropriate they are to manipulation by the powerful in society.

3.11.2 History

Although this was a postwar discovery, the Map Communication Model (MCM) has its roots in information theory developed in the telephone industry before the war began. Mathematician, inventor, and teacher Claude Shannon worked at Bell Labs after completing his Ph.D. at the Massachusetts Institute of Technology in 1940. Shannon applied mathematical theory to information and demonstrated that communication could be reduced to binary digits (bits) of positive and negative circuits. This information could be coded and transmitted across a noisy interface without losing any meaning. Once the information was received it was then decoded by the listener; the integrity of the information remained intact. In producing meaningful sounds that could be measured for quality, Shannon produced the beginning of information theory and digital communication through circuits of on and off switches.

Shannon developed his ideas more thoroughly in the 1940s at the same time that geographer and cartographer Arthur H. Robinson returned from the Second World War during which he had served as a cartographer for the military. Robinson found that cartographers were significantly limited because artists could make more effective maps than geographers. Upon returning from the war, Robinson worked to remedy this problem at Ohio State University where he was a graduate student. His *The Look of Maps* emphasizes the importance of lettering, map design, map structure, color, and technique.

Information theory helped turn the map into a medium of communicating information. Although Robinson never articulated a map model that could govern the new scientific pursuit of maps, his role in the war led to an understanding of the practical need for maps based on science not art. Robinson opened the door for others to apply Shannon's *Mathematical Theory of Communication* to the design of maps. The British geographer Christopher Board developed the first MCM in 1967 but it was cumbersome and poorly measured a map's information quality. The Czech Geographer Kolácný's 1969 version made

several key improvements to the Board's model. These versions of the MCM helped cartographers realize the problems that Robinson noted as a war cartographer and helped articulate the discipline in terms of science.

3.11.3 Map symbolization

Map Symbolization is the characters, letters, or similar graphic representations used on a map to indicate an object or characteristic in the real world.

3.11.4 Cognitive Issues

There are many cognitive issues involved in the cartographic process and symbolization (Olson, 2006). Some people may perceive certain objects differently than others, so cartographers try not to use symbols that could be easily confused. For example, red and blue are universally known to depict hot and cold.

In map-making, the principles of cognition are important since they explain why certain map symbols work (Olson, 2006). In the past, mapmakers did not care why they worked. This behaviorist view treats the human brain like a black box. Modern cartographers are curious why certain symbols are the most effective. This should help develop a theoretical basis for how brains recognize symbols and, in turn, provide a platform for creating new symbols.

3.11.5 Topographic Maps

Topographic maps show the shape of Earth's surface by using contour lines, the lines on the map that join points of equal elevation. They are among the most well-known symbols on modern maps as they are self-explanatory and accurately represent their phenomena. They make it possible to the depict height, depth, and even slope. The contour lines will be closer together or spaced apart to show the steepness of the area. If the line is spaced closer together, it means that there is a steeper slope. If they are farther apart, the area has a low slope. An area of low slope generally uses contour intervals of 10 feet or less. Areas that contain mountain or other high slope can use an interval of 100 feet (Topographic Map Symbols, 2005).

Apart from showing just contour lines, topographic maps also use a lot of map symbols to represent its features. Features are represented by using point, line, and area symbols. Individual features, such as houses, are shown as point symbols like a small dot or square. However, a cluster of houses or neighborhood can be shown as a shaded area or polygon. Areas of importance or landmarks may receive special symbols that represent what they are. For instance, a church may be symbolized as a picture of a little church or a cross or the town hall may have a special color or symbol.

3.11.6 Shape and Color of Topographic Symbols

Many of the features will be shown by straight, curved, dashed, or solid lines. They may also be colored to represent different classes of information. The typical color standard for topographic maps depicts the contours in brown, bodies of water in blue, boundaries in black, and grids and roads in red. Topographic maps may use different colors to represent area features. Most topographic maps will use green for vegetation or national parks and wildlife management areas. They will also use blue for rivers, lakes, or other bodies of water. Red may also be used to represent areas of significant importance (Topographic Map Symbols, 2005).

A map is a smaller representation of an area on the earth's surface; therefore, map symbols are used to represent real objects. Without symbols, maps would not be possible (Map Symbols). Both shapes and colors can be used for symbols on maps. A small circle may mean a point of interest, with a brown circle meaning recreation, red circle meaning services, and green circle meaning rest stop. Colors may cover larger areas of a map, such as green representing forested land and blue representing waterways. To ensure that a person can correctly read a map, a Map Legend is a key to all the symbols used on a map. It is like a dictionary so you can understand the meaning of what the map represents (Map Symbols).

3.11.7 Rules to Follow

There are certain rules to follow with map symbols. The representative symbols should always be placed on the left and defined to the right. This allows for the reader to view the symbol first, then its definition, which is customary in English dictionaries. In most cases, representative symbols should be vertically displayed and the symbols should be horizontally centered. The symbols should be vertically centered with the definitions. The definitions are supposed to be horizontally centered to the left.

3.11.8 Representing Spatial Phenomena

Symbols are used to represent geographic phenomena. Most phenomena can be represented by using pointed, line, or area symbols (Krygier & Wood, 2005). It is necessary to consider the spatial arrangement of the phenomena to determine what kind of symbolization it will require. Discrete phenomena occur at isolated points, whereas continuous phenomena occur everywhere. Both of these can also be broken down into either smooth or abrupt. For example, rainfall and taxes for states are both continuous in nature, but rainfall is smooth because it does not vary at state boundaries, leaving the tax to be considered abrupt. It is important to distinguish between the real world and the data we use to represent it. There are basically five types of spatial dimensions that are used to classify phenomena for map symbolization. Point phenomena are assumed to have no spatial extent and are said to be zero-dimensional. These uses point symbols on a map to indicate their location. An example of these would be fire hydrants or trees in a park. Linear phenomena are one-dimensional and have a length. This would include any line feature on a map like roads or sidewalks. Areal phenomena are 2-D that has both a length and a width. The best example of this would be a lake or other body of water. When volume comes into consideration, it is broken down into two types, 2 ½ dimensions and 3-D. A good example of 2 ½ D would be the elevation of a place above sea level, while 3-D being any three-dimensional objects.

3.11.9 Ranking

An important factor in map symbols is the order in which they are ranked according to their relative importance. This is known as intellectual hierarchy. The most important hierarchy is the thematic symbols and type labels that are directly related to the theme. Next comes the title, subtitle, and legend (Krygier & Wood, 2005). The map must also contain base information, such as boundaries, roads, and place names. Data sources and notes should be on all maps. Lastly, the scale, neat lines, and north arrow are the least important of the hierarchy of the map. From this we see that the symbols are the single most important thing to build a good visual hierarchy that shows a proper graphical representation. When producing a map with good visual hierarchy, thematic symbols should be graphically emphasized. A map with a visual hierarchy that is effective attracts the map user's eyes to the symbols with the most important aspects of the map first and to the symbols with the lesser importance later.

The legend of the map also contains important information and all of the thematic symbols of the map. A symbol that needs no explanation, or do not coincide with the theme of the map, are normally omitted

from the map legend. Thematic symbols directly represent the map's theme and should stand out (Map Symbols).

3.12 Choropleth Maps

Choropleth mapping is commonly used to show data for counties, states, or other enumeration units. Data collected for choropleth maps is usually grouped into separate classes based on attributes or other forms of classification. The classes are given a specific color or shading based on their values and what they are trying to portray. Choropleth maps are most effective when the data or classes change abruptly at each enumerated boundary (Slocum, McMaster, Kessler, & Howard, 2005).

A proportional symbol map is better than choropleth maps for showing raw data totals. A proportional symbols map uses symbols that are proportional to the data that they are representing with point locations. These symbols can be true points or conceptual points. True points represent real objects or the exact location of a tangible object. This could be an oil well or fire hydrant. A conceptual point represents the center of the enumeration unit, such as a corn field. The raw data on proportional symbol maps go hand in hand with the data shown on choropleth maps (Slocum, McMaster, Kessler, & Howard, 2005).

Review Questions

1. Define the map?
2. Explain the Isopleth Maps?
3. Explain the Choropleth Maps?
4. Explain the Dot Maps?

Discussion Questions

Discuss the Map projection?

Chapter 4- Statistical Methods

Learning Objectives

- To define the Central Tendency.
- To explain the Mean.
- To explain the Median.
- To describe the Mode.

4.1 Measurements of central tendencies

In statistics, a **central tendency** (or, more commonly, a **measure of central tendency**) is a central value or a typical value for a probability distribution. It is occasionally called an average or just the **center** of the distribution. The most common measures of central tendency are the arithmetic mean, the median and the mode. A central tendency can be calculated for either a finite set of values or for a theoretical distribution, such as the normal distribution. Occasionally authors use central tendency (or **centrality**), to mean "the tendency of quantitative data to cluster around some central value,". This meaning might be expected from the usual dictionary definitions of the words tendency and centrality. Those authors may judge whether the data have a strong or a weak central tendency based on the statistical dispersion, as measured by the standard deviation or something similar.

The term "central tendency" dates from the late 1920s.

The following may be applied to one-dimensional data. Depending on the circumstances, it may be appropriate to transform the data before calculating a central tendency. Examples are squaring the values or taking logarithms. Whether a transformation is appropriate and what it should be depend heavily on the data being analyzed.

- The arithmetic mean (or simply, mean) – the sum of all measurements divided by the number of observations in the data set
- Median – the middle value that separates the higher half from the lower half of the data set. The median and the mode are the only measures of central tendency that can be used for ordinal data, in which values are ranked relative to each other but are not measured absolutely.
- Mode – the most frequent value in the data set. This is the only central tendency measure that can be used with nominal data, which have purely qualitative category assignments.
- Geometric mean – the *n*th root of the product of the data values, where there are *n* of these. This measure is valid only for data that are measured absolutely on a strictly positive scale.
- Harmonic mean – the reciprocal of the arithmetic mean of the reciprocals of the data values. This measure too is valid only for data that are measured absolutely on a strictly positive scale.
- Weighted mean – an arithmetic mean that incorporates weighting of certain data elements
- Truncated mean – the arithmetic mean of the data values after a certain number or proportion of the highest and lowest data values have been discarded.
    - Interquartile mean (a type of truncated mean)
- Midrange – the arithmetic mean of the maximum and minimum values of a data set.
- Midhinge – the arithmetic mean of the two quartiles.
- Trimean – the weighted arithmetic mean of the median and two quartiles.
- Winsorized mean – an arithmetic mean in which extreme values are replaced by values closer to the median.

Any of the above may be applied to each dimension of multi-dimensional data, but the results may not be invariant to rotations of the multi-dimensional space. In addition, there is the

- Geometric median - which minimizes the sum of distances to the data points. This is the same as the median when applied to one-dimensional data, but it is not the same as taking the median of each dimension independently. It is not invariant to different rescaling of the different dimensions.

4.1.1 Solutions to variational problems

Several measures of central tendency can be characterized as solving a variational problem, in the sense of the calculus of variations, namely minimizing variation from the center. That is, given a measure of statistical dispersion, one asks for a measure of central tendency that minimizes variation: such that variation from the center is minimal among all choices of center. In a quip, "dispersion precedes location". In the sense of L^p spaces, the correspondence is:

L^p	dispersion	central tendency
L^1	average absolute deviation	median
L^2	standard deviation	mean
L^∞	maximum deviation	midrange

Thus standard deviation about the mean is lower than standard deviation about any other point, and the maximum deviation about the midrange is lower than the maximum deviation about any other point. The uniqueness of this characterization of mean follows from convex optimization. Indeed, for a given (fixed) data set *x*, the function

$$f_2(c) = \|x - c\|_2$$

represents the dispersion about a constant value *c* relative to the L^2 norm. Because the function f_2 is a strictly convex coercive function, the minimizer exists and is unique.

Note that the median in this sense is not in general unique, and in fact any point between the two central points of a discrete distribution minimizes average absolute deviation. The dispersion in the L^1 norm, given by

$$f_1(c) = \|x - c\|_1$$

is not *strictly* convex, whereas strict convexity is needed to ensure uniqueness of the minimizer. In spite of this, the minimizer is unique for the L^∞ norm.

4.2 Arithmetic mean

In mathematics and statistics, the **arithmetic mean**, or simply the mean or **average** when the context is clear, is the sum of a collection of numbers divided by the number of numbers in the collection. The collection is often a set of results of an experiment, or a set of results from a survey. The term "arithmetic mean" is preferred in some contexts in mathematics and statistics because it helps distinguish it from other means such as the geometric mean and the harmonic mean.

In addition to mathematics and statistics, the arithmetic mean is used frequently in fields such as economics, sociology, and history, and it is used in almost every academic field to some extent. For example, per capita income is the arithmetic average income of a nation's population.

While the arithmetic mean is often used to report central tendencies, it is not a robust statistic, meaning that it is greatly influenced by outliers (values that are very much larger or smaller than most of the values). Notably, for skewed distributions, such as the distribution of income for which a few people's incomes are substantially greater than most people's, the arithmetic mean may not accord with one's notion of "middle", and robust statistics such as the median may be a better description of central tendency.

In a more obscure usage, any sequence of values that form an arithmetic sequence between two numbers *x* and *y* can be called "arithmetic means between *x* and *y*."

4.2.1 **Definition**

Suppose we have a data set containing the values $a_1, \ldots, a_n$. The arithmetic mean A is defined by the formula

$$A = \frac{1}{n}\sum_{i=1}^{n} a_i.$$

If the data set is a statistical population (i.e., consists of every possible observation and not just a subset of them), then the mean of that population is called the **population mean**. If the data set is a statistical sample (a subset of the population) we call the statistic resulting from this calculation a **sample mean**.

The arithmetic mean of a variable is often denoted by a bar, for example as in $\bar{x}$ (read "x bar"), which is the mean of the n values $x_1, x_2, \ldots, x_n$.

4.2.2 **Motivating properties**

The arithmetic mean has several properties that make it useful, especially as a measure of central tendency. These include:

- If numbers $x_1, \ldots, x_n$ have mean $\bar{x}$, then $(x_1 - \bar{x}) + \cdots + (x_n - \bar{x}) = 0$. Since $x_i - \bar{x}$ is the distance from a given number to the mean, one way to interpret this property is as saying that the numbers to the left of the mean are balanced by the numbers to the right of the mean. The mean is the only single number for which the residuals (deviations from the estimate) sum to zero.
- If it is required to use a single number as a "typical" value for a set of known numbers $x_1, \ldots, x_n$, then the arithmetic mean of the numbers does this best, in the sense of minimizing the sum of squared deviations from the typical value: the sum of $(x_i - \bar{x})^2$. (It follows that the sample mean is also the best single predictor in the sense of having the lowest root mean squared error.) If the arithmetic mean of a population of numbers is desired, then the estimate of it that is unbiased is the arithmetic mean of a sample drawn from the population.

4.2.3 **Contrast with median**

The arithmetic mean may be contrasted with the median. The median is defined such that half the values are larger than, and half are smaller than, the median. If elements in the sample data, increase arithmetically, when placed in some order, then the median and arithmetic average are equal. For example, consider the data sample {1,2,3,4}. The average is 2.5, as is the median. However, when we consider a sample that cannot be arranged so as to increase arithmetically, such as {1,2,4,8,16}, the median and arithmetic average can differ significantly. In this case the arithmetic average is 6.2 and the median is 4. In general the average value can vary significantly from most values in the sample, and can be larger or smaller than most of them.

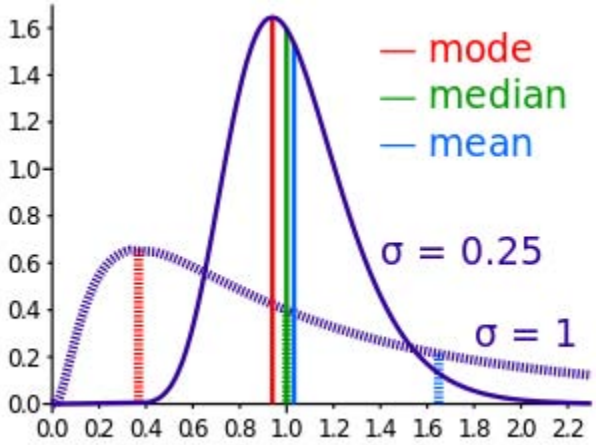
There are applications of this phenomenon in many fields. For example, since the 1980s in the United States median income has increased more slowly than the arithmetic average of income.

4.2.4 **Generalizations**

4.2.4.1 **Weighted average**

A weighted average or weighted mean is an average in which some data points count more strongly than others, in that they are given more weight in the calculation. For example, the arithmetic mean of 3 and 5 is (3+5)/2 = 4, or equivalently [(1/2) × 3] + [(1/2)× 5] = 4. In contrast, a *weighted* mean in which the first number receives twice as much weight as the second (perhaps because it is assumed to appear twice as often in the general population from which these numbers were sampled) would be calculated as [(2/3) × 3] + [(1/3) × 5] = 11/3. Here the weights, which necessarily sum to the value one, are (2/3) and (1/3), the former being twice the latter. Note that the arithmetic mean (sometimes called the "unweighted average" or "equally weighted average") can be interpreted as a special case of a weighted average in which all the weights are equal to each other (equal to 1/2 in the above example, and equal to 1/*n* in a situation with *n* numbers being averaged).

4.2.4.2 **Continuous probability distributions**



Comparison of mean, median and mode of two log-normal distributions with different skewness.

When a population of numbers, and any sample of data from it, could take on any of a continuous range of numbers, instead of for example just integers, then the probability of a number falling into one range of possible values could differ from the probability of falling into a different range of possible values, even if the lengths of both ranges are the same. In such a case the set of probabilities can be described using a

continuous probability distribution. The analog of a weighted average in this context, in which there are an infinitude of possibilities for the precise value of the variable, is called the *mean of the probability distribution*. The most widely encountered probability distribution is called the normal distribution; it has the property that all measures of its central tendency, including not just the mean but also the aforementioned median and the mode, are equal to each other. This property does not hold, however, in the cases of a great many probability distributions, such as the lognormal distribution illustrated here.

4.2.4.3 **Angles**

Particular care must be taken when using cyclic data such as phases or angles. Naïvely taking the arithmetic mean of 1° and 359° yields a result of 180°. This is incorrect for two reasons:

- Firstly, angle measurements are only defined up to an additive constant of 360° (or 2π, if measuring in radians). Thus one could as easily call these 1° and −1°, or 361° and 719°, each of which gives a different average.
- Secondly, in this situation, 0° (equivalently, 360°) is geometrically a better *average* value: there is lower dispersion about it (the points are both 1° from it, and 179° from 180°, the putative average).

In general application such an oversight will lead to the average value artificially moving towards the middle of the numerical range. A solution to this problem is to use the optimization formulation (viz., define the mean as the central point: the point about which one has the lowest dispersion), and redefine the difference as a modular distance (i.e., the distance on the circle: so the modular distance between 1° and 359° is 2°, not 358°).

4.3 **Geometric mean**

In mathematics, the **geometric mean** is a type of mean or average, which indicates the central tendency or typical value of a set of numbers by using the product of their values (as opposed to the arithmetic mean which uses their sum). The geometric mean is defined as the *n*th root (where n is the count of numbers) of the product of the numbers.

For instance, the geometric mean of two numbers, say 2 and 8, is just the square root of their product; that is $\sqrt{2 \cdot 8} = 4$. As another example, the geometric mean of the three numbers 4, 1, and 1/32 is the cube root of their product (1/8), which is 1/2; that is $\sqrt[3]{4 \cdot 1 \cdot 1/32} = 1/2$.

A geometric mean is often used when comparing different items – finding a single "figure of merit" for these items – when each item has multiple properties that have different numeric ranges. For example, the geometric mean can give a meaningful "average" to compare two companies which are each rated at 0 to 5 for their environmental sustainability, and are rated at 0 to 100 for their financial viability. If an arithmetic mean was used instead of a geometric mean, the financial viability is given more weight because its numeric range is larger- so a small percentage change in the financial rating (e.g. going from 80 to 90) makes a much larger difference in the arithmetic mean than a large percentage change in environmental sustainability (e.g. going from 2 to 5). The use of a geometric mean "normalizes" the ranges being averaged, so that no range dominates the weighting, and a given percentage change in any of the properties has the same effect on the geometric mean. So, a 20% change in environmental sustainability from 4 to 4.8 has the same effect on the geometric mean as a 20% change in financial viability from 60 to 72.

The geometric mean can be understood in terms of geometry. The geometric mean of two numbers, $a$and b, is the length of one side of a square whose area is equal to the area of a rectangle with sides of lengths $a$and b. Similarly, the geometric mean of three numbers, a, b, and c, is the length of one side of a cube whose volume is the same as that of a cuboid with sides whose lengths are equal to the three given numbers.

The geometric mean applies only to positive numbers. It is also often used for a set of numbers whose values are meant to be multiplied together or are exponential in nature, such as data on the growth of the human population or interest rates of a financial investment.

The geometric mean is also one of the three classical Pythagorean means, together with the aforementioned arithmetic mean and the harmonic mean. For all positive data sets containing at least one pair of unequal values, the harmonic mean is always the least of the three means, while the arithmetic mean is always the greatest of the three and the geometric mean is always in between.

4.3.1 **Calculation**

The geometric mean of a data set $\{a_1, a_2, \ldots, a_n\}$is given by:

$$\left(\prod_{i=1}^{n} a_i\right)^{1/n} = \sqrt[n]{a_1 a_2 \cdots a_n}.$$

The geometric mean of a data set is less than the data set's arithmetic mean unless all members of the data set are equal, in which case the geometric and arithmetic means are equal. This allows the definition of the arithmetic-geometric mean, a mixture of the two which always lies in between.

The geometric mean is also the **arithmetic-harmonic mean** in the sense that if two sequences (a_n) and (h_n) are defined:

$$a_{n+1} = \frac{a_n + h_n}{2}, \quad a_0 = x$$

and

$$h_{n+1} = \frac{2}{\frac{1}{a_n} + \frac{1}{h_n}}, \quad h_0 = y$$

where $h_{n+1}$is the harmonic mean of the previous values of the two sequences, then $a_n$and $h_n$will converge to the geometric mean of $x$and y.

This can be seen easily from the fact that the sequences do converge to a common limit (which can be shown by Bolzano–Weierstrass theorem) and the fact that the geometric mean is preserved:

$$\sqrt{a_i h_i} = \sqrt{\frac{a_i + h_i}{\frac{a_i + h_i}{h_i a_i}}} = \sqrt{\frac{a_i + h_i}{\frac{1}{a_i} + \frac{1}{h_i}}} = \sqrt{a_{i+1} h_{i+1}}$$

Replacing the arithmetic and harmonic mean by a pair of generalized means of opposite, finite exponents yields the same result.

4.3.2 **Relationship with arithmetic mean of logarithms**

By using logarithmic identities to transform the formula, the multiplications can be expressed as a sum and the power as a multiplication.

$$\left(\prod_{i=1}^{n} a_i\right)^{1/n} = \exp\left[\frac{1}{n}\sum_{i=1}^{n} \ln a_i\right]$$

This is sometimes called the **log-average**. It simply computes the arithmetic mean of the logarithm-transformed values of a_i(i.e., the arithmetic mean on the log scale) and then using the exponentiation to return the computation to the original scale, i.e., It is the generalized f-mean with $f(x) = \log x$. For example, the geometric mean of 2 and 8 can be calculated as:

$$b^{(\log_b(2)+\log_b(8))/2} = 4,$$

where $b$is any base of a logarithm (commonly 2, $e$or 10).

The right-hand side formula above is generally the preferred alternative for implementation in computer languages: overflows or underflows are less likely to happen compared to calculating the product of a set of numbers due to taking logarithms.

4.3.3 **Relationship with arithmetic mean and mean-preserving spread**

If a set of non-identical numbers is subjected to a mean-preserving spread — that is, two or more elements of the set are "spread apart" from each other while leaving the arithmetic mean unchanged — then the geometric mean always decreases.

4.3.4 **Computation in constant time**

In cases where the geometric mean is being used to determine the average growth rate of some quantity, and the initial and final values $a_0$and $a_n$of that quantity are known, the product of the measured growth rate at every step need not be taken. Instead, the geometric mean is simply

$$\left(\frac{a_n}{a_0}\right)^{\frac{1}{n}},$$

where $n$is the number of steps from the initial to final state.

If the values are $a_0, \cdots, a_n$, then the growth rate between measurement $a_k$and $a_{k+1}$is a_{k+1}/a_k. The geometric mean of these growth rates is just

$$\left(\frac{a_1}{a_0}\frac{a_2}{a_1}\cdots\frac{a_n}{a_{n-1}}\right)^{\frac{1}{n}} = \left(\frac{a_n}{a_0}\right)^{\frac{1}{n}}$$

4.3.5 **Properties**

The fundamental property of the geometric mean, which can be proven to be false for any other mean, is

$$GM\left(\frac{X_i}{Y_i}\right) = \frac{GM(X_i)}{GM(Y_i)}$$

This makes the geometric mean the only correct mean when averaging *normalized* results, that is results that are presented as ratios to reference values. This is the case when presenting computer performance with respect to a reference computer, or when computing a single average index from several heterogeneous sources (for example life expectancy, education years and infant mortality). In this scenario, using the arithmetic or harmonic mean would change the ranking of the results depending on what is used as a reference. For example, take the following comparison of execution time of computer programs:

	Computer A	**Computer B**	**Computer C**
Program 1	1	10	20
Program 2	1000	100	20
Arithmetic mean	500.5	55	**20**
Geometric mean	31.622 . . .	31.622 . . .	**20**

The arithmetic and geometric means "agree" that computer C is the fastest. However, by presenting appropriately normalized values *and* using the arithmetic mean, we can show either of the other two computers to be the fastest. Normalizing by A's result gives A as the fastest computer according to the arithmetic mean:

	Computer A	**Computer B**	**Computer C**
Program 1	1	10	20
Program 2	1	0.1	0.02
Arithmetic mean	**1**	5.05	10.01
Geometric mean	1	1	**0.632 . . .**

while normalizing by B's result gives B as the fastest computer according to the arithmetic mean:

	Computer A	**Computer B**	**Computer C**
Program 1	0.1	1	2
Program 2	10	1	0.2
Arithmetic mean	5.05	**1**	1.1

Geometric mean	1	1	**0.632**

In all cases, the ranking given by the geometric mean stays the same as the one obtained with unnormalized values.

4.3.6 **Applications**

4.3.6.1 **Proportional growth**

The geometric mean is more appropriate than the arithmetic mean for describing proportional growth, both exponential growth (constant proportional growth) and varying growth; in business the geometric mean of growth rates is known as the compound annual growth rate (CAGR). The geometric mean of growth over periods yields the equivalent constant growth rate that would yield the same final amount.

Suppose an orange tree yields 100 oranges one year and then 180, 210 and 300 the following years, so the growth is 80%, 16.6666% and 42.8571% for each year respectively. Using the arithmetic mean calculates a (linear) average growth of 46.5079% (80% + 16.6666% + 42.8571% divided by 3). However, if we start with 100 oranges and let it grow 46.5079% each year, the result is 314 oranges, not 300, so the linear average *over*-states the year-on-year growth.

Instead, we can use the geometric mean. Growing with 80% corresponds to multiplying with 1.80, so we take the geometric mean of 1.80, 1.166666 and 1.428571, i.e. $\sqrt[3]{1.80 \times 1.166666 \times 1.428571} = 1.442249$; thus the "average" growth per year is 44.2249%. If we start with 100 oranges and let the number grow with 44.2249% each year, the result is 300 oranges.

4.3.6.2 **Applications in the social sciences**

Although the geometric mean has been relatively rare in computing social statistics, starting from 2010 the United Nations Human Development Index did switch to this mode of calculation, on the grounds that it better reflected the non-substitutable nature of the statistics being compiled and compared:

> *The geometric mean decreases the level of substitutability between dimensions [being compared] and at the same time ensures that a 1 percent decline in say life expectancy at birth has the same impact on the HDI as a 1 percent decline in education or income. Thus, as a basis for comparisons of achievements, this method is also more respectful of the intrinsic differences across the dimensions than a simple average.*

Note that not all values used to compute the HDI are normalized; some of them instead have the form $(X - X_{\mathbf{min}})/(X_{\mathbf{norm}} - X_{\mathbf{min}})$. This makes the choice of the geometric mean less obvious than one would expect from the "Properties" section above.

4.3.6.3 **Aspect ratios**

The geometric mean has been used in choosing a compromise aspect ratio in film and video: given two aspect ratios, the geometric mean of them provides a compromise between them, distorting or cropping both in some sense equally. Concretely, two equal area rectangles (with the same center and parallel sides) of different aspect ratios intersect in a rectangle whose aspect ratio is the geometric mean, and their hull (smallest rectangle which contains both of them) likewise have an aspect ratio their geometric mean.

In the choice of 16:9 aspect ratio by the SMPTE, balancing 2.35 and 4:3, the geometric mean is $\sqrt{2.35 \times \frac{4}{3}} \approx 1.7701$, and thus $16 : 9 = 1.77\overline{7}...$ was chosen. This was discovered empirically by Kerns Powers, who cut out rectangles with equal areas and shaped them to match each of the popular aspect ratios. When overlapped with their center points aligned, he found that all of those aspect ratio rectangles fit within an outer rectangle with an aspect ratio of 1.77:1 and all of them also covered a smaller common inner rectangle with the same aspect ratio 1.77:1. The value found by Powers is exactly the geometric mean of the extreme aspect ratios, 4:3 (1.33:1) and CinemaScope (2.35:1), which is coincidentally close to $16 : 9 (1.77\overline{7} : 1)$. Note that the intermediate ratios have no effect on the result, only the two extreme ratios.

Applying the same geometric mean technique to 16:9 and 4:3 approximately yields the 14:9 ($1.55\overline{5}...$) aspect ratio, which is likewise used as a compromise between these ratios. In this case 14:9 is exactly the *arithmetic mean* of $16 : 9$ and $4 : 3 = 12 : 9$, since 14 is the average of 16 and 12, while the precise *geometric mean* is $\sqrt{\frac{16}{9} \times \frac{4}{3}} \approx 1.5396 \approx 13.8 : 9,$ but the two different *means*, arithmetic and geometric, are approximately equal because both numbers are sufficiently close to each other (a difference of less than 2%).

4.4 **Harmonic mean**

In mathematics, the **harmonic mean** (sometimes called the **subcontrary mean**) is one of several kinds of mean and hence one of several kinds of average. Typically, it is appropriate for situations when the average of rates is desired.

It is the special case (M^{-1}) of the power mean. As it tends strongly toward the least elements of the list, it may (compared to the arithmetic mean) mitigate the influence of large outliers and increase the influence of small values.

The harmonic mean is one of the Pythagorean means, along with the arithmetic mean and the geometric mean, and is no greater than either of them.

4.4.1 **Definition**

4.4.1.1 **Discrete distribution**

The harmonic mean H of the positive real numbers $x_1, x_2, ..., x_n$ is defined to be the reciprocal of the arithmetic mean of the reciprocals of $x_1, x_2, ..., x_n$:

$$H = \left(\frac{1}{n} \cdot \sum_{i=1}^{n} x_i^{-1} \right)^{-1} = \frac{1}{\frac{1}{n} \cdot \left(\frac{1}{x_1} + \frac{1}{x_2} + \cdots + \frac{1}{x_n} \right)} = \frac{n}{\frac{1}{x_1} + \frac{1}{x_2} + \cdots + \frac{1}{x_n}}.$$

Example

The harmonic mean of 1, 2, and 4 is

$$\frac{3}{\frac{1}{1}+\frac{1}{2}+\frac{1}{4}} = \frac{1}{\frac{1}{3}(\frac{1}{1}+\frac{1}{2}+\frac{1}{4})} = \frac{12}{7} = 1.\overline{714285}.$$

4.4.1.2 **Continuous distribution**

For a continuous distribution the harmonic mean is

$$H = \frac{b-a}{\int_a^b \frac{1}{f(x)}\,dx}.$$

4.4.2 **Weighted harmonic mean**

If a set of weights $w_1, \ldots, w_n$ is associated with the dataset $x_1, \ldots, x_n$, the **weighted harmonic mean** is defined by

$$\frac{\sum_{i=1}^n w_i}{\sum_{i=1}^n \frac{w_i}{x_i}}.$$

The harmonic mean is a special case where all of the weights are equal to 1. It is also equivalent to any weighted harmonic mean where all weights are equal.

4.4.3 **Recursive calculation**

It is possible to recursively calculate the harmonic mean (*H*) of *n* variates. This method may be of use in computations.

$$H(x_1, x_2, x_3, \ldots) = \frac{n}{\sum \frac{1}{x_i}} = \left(\frac{1}{n}x_1^{-1} + \frac{n-1}{n}H(x_2, x_3, \ldots)^{-1}\right)^{-1}$$

4.4.4 **Harmonic mean of two numbers**

For the special case of just two numbers x_1 and x_2, the harmonic mean can be written

$$H = \frac{2x_1x_2}{x_1+x_2}.$$

In this special case, the harmonic mean is related to the arithmetic mean $A = \frac{x_1+x_2}{2}$ and the geometric mean $G = \sqrt{x_1x_2}$, by

$$H = \frac{G^2}{A}.$$

So

$$G = \sqrt{AH}$$

meaning the two numbers' geometric mean equals the geometric mean of their arithmetic and harmonic means.

As noted above this relationship between the three Pythagorean means is not limited to *n* equals 1 or 2; there is a relationship for all *n*. However, for *n* = 1 all means are equal and for *n* = 2 we have the above relationship between the means. For arbitrary $n \geq 2$ we may generalize this formula, as noted above, by interpreting the third equation for the harmonic mean differently. The generalized relationship was already explained above. The third equation also works for *n* = 1. That is, it predicts the equivalence between the harmonic and geometric means but it falls short by not predicting the equivalence between the harmonic and arithmetic means.

The general formula, which can be derived from the third formula for the harmonic mean by the reinterpretation as explained in relationship with other means, is

$$H(x_1, \ldots, x_n) = \frac{(G(x_1, \ldots, x_n))^n}{A(x_2 x_3 \cdots x_n, x_1 x_3 \cdots x_n, \ldots, x_1 x_2 \cdots x_{n-1})} = \frac{(G(x_1, .}{A\left(\frac{\prod_{i=1}^{n} x_i}{x_1}, \frac{\prod_{i=1}^{n}}{x_2}\right.}$$

For *n* = 2,

$$H(x_1, x_2) = \frac{(G(x_1, x_2))^2}{A(x_2, x_1)} = \frac{(G(x_2, x_1))^2}{A(x_2, x_1)}$$

where we used the fact that the arithmetic mean evaluates to the same number independent of the order of the terms. This equation can be reduced to the original equation if we reinterpret this result in terms of the operators themselves. If we do this we get the symbolic equation

$$H = \frac{G^2}{A}$$

because each function was evaluated at

$$(x_1, x_2).$$

4.4.5 **Relationship with other means**

If a set of non-identical numbers is subjected to a mean-preserving spread — that is, two or more elements of the set are "spread apart" from each other while leaving the arithmetic mean unchanged — then the harmonic mean always decreases.

Let *r* be a non zero real number and let the r^{th} power mean (M^r) of a series of real variables (a_1, a_2, a_3, ...) be defined as

$$M^r(a_1, a_2, a_3, \dots) = \left(\frac{1}{n}\sum (a_i)^r\right)^{\frac{1}{r}}.$$

For r = -1, 1 and 2 we have the harmonic, the arithmetic and the quadratic means respectively. Define $M^r(.)$ for r = 0, -∞ and +∞ to be the geometric mean, the minimum of the variates and the maximum of the variates respectively. Then for any two real numbers s and t such that $s < t$ we have

$$M^s(a_1, a_2, a_3, \dots) \leq M^t(a_1, a_2, a_3, \dots).$$

with equality only if all the a_i are equal.

Let R be the quadratic mean (or root mean square). Then

$$\frac{2R + H}{3} \leq A.$$

4.4.6 **Inequalities**

For a set of positive real variables lying within the interval [m, M] it has been shown that

$$A - H \geq \frac{s^2}{2M}$$

where A is the arithmetic mean, H is the harmonic mean, M is the maximum of the interval and s^2 is the variance of the set.

Several other inequalities are also known:

$$\frac{m(A-m)(A-H)}{H-m} \leq s^2 \leq \frac{M(A-H)(M-A)}{M-H}$$
$$\frac{(M-s)^2}{M(M-2s)} \leq \frac{A}{H} \leq \frac{(m+s)^2}{m(m+2s)}$$
$$\frac{(M-m)s^2}{M(M-m)-s^2} \leq A - H \leq \frac{(M-m)s^2}{m(M-m)+s^2}$$

4.4.7 **Examples**

Geometry

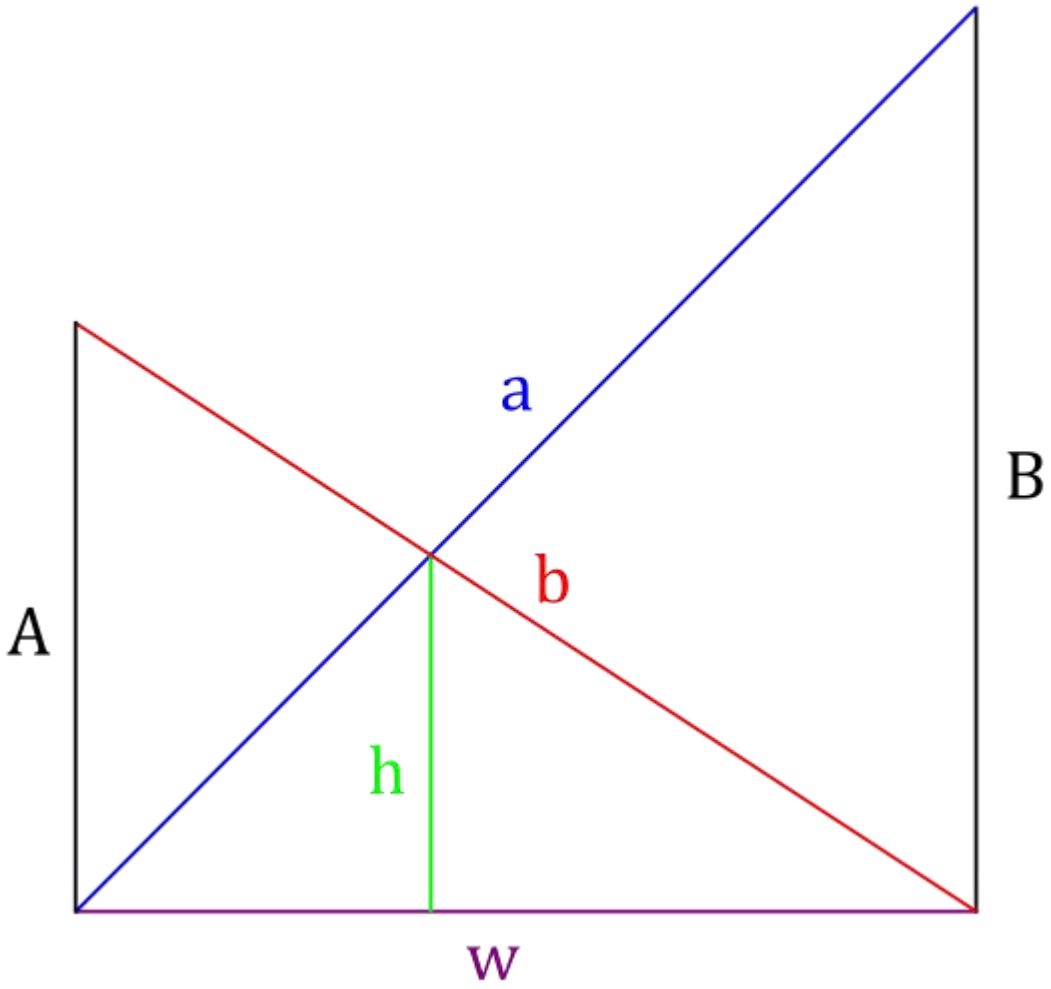
In any triangle, the radius of the incircle is one-third the harmonic mean of the altitudes.

For any point P on the minor arc BC of the circumcircle of an equilateral triangle ABC, with distances q and t from B and C respectively, and with the intersection of PA and BC being at a distance y from point P, we have that y is half the harmonic mean of q and t.

In a right triangle with legs a and b and altitude h from the hypotenuse to the right angle, h^2 is half the harmonic mean of a^2 and b^2.

Let t and s ($t > s$) be the sides of the two inscribed squares in a right triangle with hypotenuse c. Then s^2 equals half the harmonic mean of c^2 and t^2.

Let a trapezoid have vertices A, B, C, and D in sequence and have parallel sides AB and CD. Let E be the intersection of the diagonals, and let F be on side DA and G be on side BC such that FEG is parallel to AB and CD. Then FG is the harmonic mean of AB and DC. (This is provable using similar triangles.)



Crossed ladders. h is half the harmonic mean of A and B

In the crossed ladders problem, two ladders lie oppositely across an alley, each with feet at the base of one side wall, with one leaning against a wall at height A and the other leaning against the opposite wall at height B, as shown. The ladders cross at a height of h above the alley floor. Then h is half the harmonic mean of A and B. This result still holds if the walls are slanted but still parallel and the "heights" A, B, and h are measured as distances from the floor along lines parallel to the walls.

In an ellipse, the semi-latus rectum (the distance from a focus to the ellipse along a line parallel to the minor axis) is the harmonic mean of the maximum and minimum distances of the ellipse from a focus.

4.4.8 **Trigonometry**

In the case of the double-angle tangent identity, if the tangent of an angle *A* is given as *a* / *b* then the tangent of 2*A* is the product of

- (1) The harmonic mean of the numerator and denominator of tan *A*; and
- (2) The reciprocal of the denominator less the numerator of tan *A*.

In symbols if *a* and *b* are real numbers and

$$\tan A = \frac{a}{b}$$

the double angle formula for the tangent can be written as

$$\tan 2A = H(a,b) \cdot \frac{1}{b-a} = \frac{2ab}{a+b} \cdot \frac{1}{b-a}$$

where *H*(*a*, *b*) is the harmonic mean of *a* and *b*.

Example

Let

$$\tan A = \frac{3}{7}$$

The harmonic mean of 3 and 7 is

$$H(3,7) = \frac{42}{10} = 4.2$$

The most familiar form of the double angle formula is

$$\tan 2A = \frac{2 \cdot \frac{3}{7}}{1 - (\frac{3}{7})^2} = \frac{21}{20};$$

The double angle formula can also be written as

$$\frac{2 \cdot 3 \cdot 7}{3+7} \cdot \frac{1}{7-3} = \frac{42}{10} \cdot \frac{1}{4} = \frac{21}{20} = 1.05$$

4.4.9 **Algebra**

The harmonic mean also features in elementary algebra when considering problems of working in parallel.

For example, if a gas powered pump can drain a pool in 4 hours and a battery powered pump can drain the same pool in 6 hours, then it will take both pumps

$$(6 \cdot 4)/(6+4) = \frac{1}{2}H(4,6) = 2.4$$

hours to drain the pool working together.

Another example involves calculating the average speed for a number of fixed-distance trips. For example, if the speed for going from point *A* to *B* was 60 km/h, and the speed for returning from *B* to *A* was 40 km/h, then the average speed is given by

$$\frac{2}{1/60 + 1/40} = 48.$$

4.5 Weighted arithmetic mean

The **weighted mean** is similar to an arithmetic mean (the most common type of average), where instead of each of the data points contributing equally to the final average, some data points contribute more than others. The notion of weighted mean plays a role in descriptive statistics and also occurs in a more general form in several other areas of mathematics.

If all the weights are equal, then the weighted mean is the same as the arithmetic mean. While weighted means generally behave in a similar fashion to arithmetic means, they do have a few counterintuitive properties, as captured for instance in Simpson's paradox.

4.5.1 **Basic example**

Given two school classes, one with 20 students, and one with 30 students, the grades in each class on a test were:

> Morning class = 62, 67, 71, 74, 76, 77, 78, 79, 79, 80, 80, 81, 81, 82, 83, 84, 86, 89, 93, 98
> Afternoon class = 81, 82, 83, 84, 85, 86, 87, 87, 88, 88, 89, 89, 89, 90, 90, 90, 90, 91, 91, 91, 92, 92, 93, 93, 94, 95, 96, 97, 98, 99

The straight average for the morning class is 80 and the straight average of the afternoon class is 90. The straight average of 80 and 90 is 85, the mean of the two class means. However, this does not account for the difference in number of students in each class (20 versus 30); hence the value of 85 does not reflect

the average student grade (independent of class). The average student grade can be obtained by averaging all the grades, without regard to classes (add all the grades up and divide by the total number of students):

$$\bar{x} = \frac{4300}{50} = 86.$$

Or, this can be accomplished by weighting the class means by the number of students in each class (using a weighted mean of the class means):

$$\bar{x} = \frac{(20 \times 80) + (30 \times 90)}{20 + 30} = 86.$$

Thus, the weighted mean makes it possible to find the average student grade in the case where only the class means and the number of students in each class are available.

4.5.2 **Convex combination example**

Since only the *relative* weights are relevant, any weighted mean can be expressed using coefficients that sum to one. Such a linear combination is called a convex combination.

Using the previous example, we would get the following:

$$\frac{20}{20 + 30} = 0.4$$
$$\frac{30}{20 + 30} = 0.6$$
$$\bar{x} = \frac{(0.4 \times 80) + (0.6 \times 90)}{0.4 + 0.6} = 86.$$

4.5.3 **Mathematical definition**

Formally, the weighted mean of a non-empty set of data

$$\{x_1, x_2, \ldots, x_n\},$$

with non-negative weights

$$\{w_1, w_2, \ldots, w_n\},$$

is the quantity

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

which means:

$$\bar{x} = \frac{w_1 x_1 + w_2 x_2 + \cdots + w_n x_n}{w_1 + w_2 + \cdots + w_n}.$$

Therefore data elements with a high weight contribute more to the weighted mean than do elements with a low weight. The weights cannot be negative. Some may be zero, but not all of them (since division by zero is not allowed).

The formulas are simplified when the weights are normalized such that they sum up to 1, i.e. $\sum_{i=1}^{n} w_i = 1$. For such normalized weights the weighted mean is simply $\bar{x} = \sum_{i=1}^{n} w_i x_i$.

Note that one can always normalize the weights by making the following transformation on the weights $w_i' = \frac{w_i}{\sum_{j=1}^{n} w_j}$. Using the normalized weight yields the same results as when using the original weights. Indeed,

$$\bar{x} = \sum_{i=1}^{n} w_i' x_i = \sum_{i=1}^{n} \frac{w_i}{\sum_{j=1}^{n} w_j} x_i = \frac{\sum_{i=1}^{n} w_i x_i}{\sum_{j=1}^{n} w_j} = \frac{\sum_{i=1}^{n} w_i x_i}{\sum_{i=1}^{n} w_i}.$$

The common mean $\frac{1}{n}\sum_{i=1}^{n} x_i$ is a special case of the weighted mean where all data have equal weights, $w_i = w$. When the weights are normalized then $w_i' = \frac{1}{n}$.

4.5.4 **Statistical properties**

The weighted sample mean, $\bar{x}$, with normalized weights (weights summing to one) is itself a random variable. Its expected value and standard deviation are related to the expected values and standard deviations of the observations as follows,

If the observations have expected values

$$E(x_i) = \bar{x}_i,$$

then the weighted sample mean has expectation

$$E(\bar{x}) = \sum_{i=1}^{n} w_i \bar{x}_i.$$

Particularly, if the expectations of all observations are equal, $\bar{x}_i = c$, then the expectation of the weighted sample mean will be the same,

$$E(\bar{x}) = c.$$

For uncorrelated observations with standard deviations σ_i, the weighted sample mean has standard deviation

$$\sigma(\bar{x}) = \sqrt{\sum_{i=1}^{n} w_i^2 \sigma_i^2}.$$

Consequently, when the standard deviations of all observations are equal, $\sigma_i = d$, the weighted sample mean will have standard deviation $\sigma(\bar{x}) = d\sqrt{V_2}$. Here V_2 is the quantity

$$V_2 = \sum_{i=1}^{n} w_i^2,$$

such that $1/n \leq V_2 \leq 1$. It attains its minimum value for equal weights, and its maximum when all weights except one are zero. In the former case we have $\sigma(\bar{x}) = d/\sqrt{n}$, which is related to the central limit theorem.

Note that due to the fact that one can always transform non-normalized weights to normalized weights all formula in this section can be adapted to non-normalized weights by replacing all w_i by $w_i' = \frac{w_i}{\sum_{i=1}^{n} w_i}$.

4.5.5 **Dealing with variance**

For the weighted mean of a list of data for which each element x_i comes from a different probability distribution with known variance ${\sigma_i}^2$, one possible choice for the weights is given by:

$$w_i = \frac{1}{\sigma_i^2}.$$

The weighted mean in this case is:

$$\bar{x} = \frac{\sum_{i=1}^{n}(x_i w_i)}{\sum_{i=1}^{n} w_i},$$

and the variance of the weighted mean is:

$$\sigma_{\bar{x}}^2 = \frac{1}{\sum_{i=1}^{n} w_i},$$

which reduces to $\sigma_{\bar{x}}^2 = \frac{{\sigma_0}^2}{n}$, when all $\sigma_i = \sigma_0$.

The significance of this choice is that this weighted mean is the maximum likelihood estimator of the mean of the probability distributions under the assumption that they are independent and normally distributed with the same mean.

4.5.6 **Correcting for over- or under-dispersion**

Weighted means are typically used to find the weighted mean of experimental data, rather than theoretically generated data. In this case, there will be some error in the variance of each data point. Typically experimental errors may be underestimated due to the experimenter not taking into account all sources of error in calculating the variance of each data point. In this event, the variance in the weighted mean must be corrected to account for the fact that χ^2 is too large. The correction that must be made is

$$\sigma_{\bar{x}}^2 \to \sigma_{\bar{x}}^2 \chi_\nu^2$$

where χ_ν^2 is χ^2 divided by the number of degrees of freedom, in this case $n - 1$. This gives the variance in the weighted mean as:

$$\sigma_{\bar{x}}^2 = \frac{1}{\sum_{i=1}^{n} 1/{\sigma_i}^2} \times \frac{1}{(n-1)} \sum_{i=1}^{n} \frac{(x_i - \bar{x})^2}{\sigma_i^2};$$

when all data variances are equal, $\sigma_i = \sigma_0$, they cancel out in the weighted mean variance, $\sigma_{\bar{x}}^2$, which then reduces to the standard error of the mean (squared), $\sigma_{\bar{x}}^2 = \sigma^2/n$, in terms of the sample standard deviation (squared), $\sigma^2 = \sum_{i=1}^{n} (x_i - \bar{x})^2/(n-1)$.

4.5.7 **Weighted sample variance**

Typically when a mean is calculated it is important to know the variance and standard deviation about that mean. When a weighted mean μ^* is used, the variance of the weighted sample is different from the variance of the unweighted sample. The *biased* weighted sample variance is defined similarly to the normal *biased* sample variance:

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

$$\sigma_{\text{weighted}}^2 = \frac{\sum_{i=1}^{N} w_i (x_i - \mu^*)^2}{V_1}$$

$$V_1 = \sum_{i=1}^{n} w_i$$

where $i=1$, which is 1 for normalized weights.

For small samples, it is customary to use an unbiased estimator for the population variance. In normal unweighted samples, the *N* in the denominator (corresponding to the sample size) is changed to $N - 1$. While this is simple in unweighted samples, it is not straightforward when the sample is weighted.

If each x_i is drawn from a Gaussian distribution with variance $1/w_i$, the unbiased estimator of a weighted population variance is given by:

$$s^2 = \frac{V_1}{V_1^2 - V_2} \sum_{i=1}^{N} w_i \left(x_i - \mu^*\right)^2,$$

$$V_2 = \sum_{i=1}^{n} w_i^2$$

where as introduced previously.

Note: If the weights are not integral frequencies (for instance, if they have been standardized to sum to 1 or if they represent the variance of each observation's measurement) as in this case, then all information is lost about the total sample size n, whence it is not possible to use an unbiased estimator because it is impossible to estimate the Bessel correction factor $\frac{n}{(n-1)}$.

The degrees of freedom of the weighted, unbiased sample variance vary accordingly from $N - 1$ down to 0.

The standard deviation is simply the square root of the variance above.

If all of the x_i are drawn from the same distribution and the integer weights w_i indicate the number of occurrences ("repeat") of an observation in the sample, then the unbiased estimator of the weighted population variance is given by

$$s^2 = \frac{1}{V_1 - 1} \sum_{i=1}^{N} w_i \left(x_i - \mu^*\right)^2 = \frac{1}{\sum_{i=1}^{n} w_i - 1} \sum_{i=1}^{N} w_i \left(x_i - \mu^*\right)^2,$$

If all x_i are unique, then N counts the number of unique values, and V_1 counts the number of samples.

For example, if values $\{2, 2, 4, 5, 5, 5\}$ are drawn from the same distribution, then we can treat this set as an unweighted sample, or we can treat it as the weighted sample $\{2, 4, 5\}$ with corresponding weights $\{2, 1, 3\}$, and we should get the same results.

As a side note, other approaches have been described to compute the weighted sample variance.

4.5.8 **Weighted sample covariance**

In a weighted sample, each row vector $\mathbf{x}_i$ (each set of single observations on each of the *K* random variables) is assigned a weight $w_i \geq 0$. Without loss of generality, assume that the weights are normalized:

$$\sum_{i=1}^{N} w_i = 1.$$

If they are not, divide the weights by their sum:

$$w_i' = \frac{w_i}{\sum_{i=1}^{N} w_i}$$

Then the weighted mean vector μ^* is given by

$$\mu^* = \sum_{i=1}^{N} w_i \mathbf{x}_i.$$

(if the weights are not normalized, an equivalent formula to compute the weighted mean is:)

$$\mu^* = \frac{\sum_{i=1}^{N} w_i \mathbf{x}_i}{\sum_{i=1}^{N} w_i}.$$

and the unbiased weighted covariance matrix $\mathbf{\Sigma}$is

$$\Sigma = \frac{\sum_{i=1}^{N} w_i}{\left(\sum_{i=1}^{N} w_i\right)^2 - \sum_{i=1}^{N} w_i^2} \sum_{i=1}^{N} w_i \left(\mathbf{x}_i - \mu^*\right)^T \left(\mathbf{x}_i - \mu^*\right).$$

If all weights are the same, with $w_i = 1/N$, then the weighted mean and covariance reduce to the sample mean and covariance above.

Alternatively, if each weight $w_i \geq 0$assigns a number of occurrences for one observation value, so $\mathbf{x}_i$ (sometimes called the number of "repeats") and is **unnormalized** so that $\sum_{i=1}^{N} w_i = N^*$with N^* being the sample size (total number of observations), then the weighted sample covariance matrix is given by:

$$\Sigma = \frac{1}{\sum_{i=1}^{N} w_i} \sum_{i=1}^{N} w_i \left(x_i - \mu^*\right)^T \left(x_i - \mu^*\right),$$

and the unbiased weighted sample covariance matrix is given by applying the Bessel correction (since $\sum_{i=1}^{N} w_i = N^*$ which is the real sample size):

$$\Sigma = \frac{1}{\sum_{i=1}^{N} w_i - 1} \sum_{i=1}^{N} w_i \left(x_i - \mu^*\right)^T \left(x_i - \mu^*\right).$$

4.5.9 **Vector-valued estimates**

The above generalizes easily to the case of taking the mean of vector-valued estimates. For example, estimates of position on a plane may have less certainty in one direction than another. As in the scalar case, the weighted mean of multiple estimates can provide a maximum likelihood estimate. We simply replace σ^2 by the covariance matrix:

$$W_i = \Sigma_i^{-1}.$$

The weighted mean in this case is:

$$\bar{\mathbf{x}} = \left(\sum_{i=1}^{n} \Sigma_i^{-1}\right)^{-1} \left(\sum_{i=1}^{n} \Sigma_i^{-1} \mathbf{x}_i\right),$$

and the covariance of the weighted mean is:

$$\Sigma_{\bar{\mathbf{x}}} = \left(\sum_{i=1}^{n} \Sigma_i^{-1}\right)^{-1},$$

For example, consider the weighted mean of the point [1 0] with high variance in the second component and [0 1] with high variance in the first component. Then

$$\begin{aligned} \mathbf{x}_1 &:= [10]^\top, & \Sigma_1 &:= \begin{bmatrix} 1 & 0 \\ 0 & 100 \end{bmatrix} \\ \mathbf{x}_2 &:= [01]^\top, & \Sigma_2 &:= \begin{bmatrix} 100 & 0 \\ 0 & 1 \end{bmatrix} \end{aligned}$$

then the weighted mean is:

$$\begin{aligned} \bar{\mathbf{x}} &= \left(\Sigma_1^{-1} + \Sigma_2^{-1}\right)^{-1} \left(\Sigma_1^{-1}\mathbf{x}_1 + \Sigma_2^{-1}\mathbf{x}_2\right) \\ &= \begin{bmatrix} 0.9901 & 0 \\ 0 & 0.9901 \end{bmatrix} \begin{bmatrix} 1 \\ 1 \end{bmatrix} = \begin{bmatrix} 0.9901 \\ 0.9901 \end{bmatrix} \end{aligned}$$

which makes sense: the [1 0] estimate is "compliant" in the second component and the [0 1] estimate is compliant in the first component, so the weighted mean is nearly [1 1].

4.5.10 **Accounting for correlations**

In the general case, suppose that $\mathbf{X} = [x_1, \ldots, x_n]$, $\mathbf{C}$is the covariance matrix relating the quantities x_i, $\bar{x}$is the common mean to be estimated, and $\mathbf{W}$is the design matrix [1, ..., 1] (of length n). The Gauss–Markov theorem states that the estimate of the mean having minimum variance is given by:

$$\sigma_{\bar{x}}^2 = (\mathbf{W}^T \mathbf{C}^{-1} \mathbf{W})^{-1},$$

and

$$\bar{x} = \sigma_{\bar{x}}^2 (\mathbf{W}^T \mathbf{C}^{-1} \mathbf{X}).$$

4.5.11 **Decreasing strength of interactions**

Consider the time series of an independent variable $x$and a dependent variable y, with $n$observations sampled at discrete times t_i. In many common situations, the value of $y$at time $t_i$depends not only on $x_i$but also on its past values. Commonly, the strength of this dependence decreases as the separation of observations in time increases. To model this situation, one may replace the independent variable by its sliding mean $z$for a window size m.

$$z_k = \sum_{i=1}^{m} w_i x_{k+1-i}.$$

Range weighted mean interpretation

Range (1–5)	**Weighted mean equivalence**
3.34–5.00	Strong
1.67–3.33	Satisfactory
0.00–1.66	Weak

4.5.12 **Exponentially decreasing weights**

In the scenario described in the previous section, most frequently the decrease in interaction strength obeys a negative exponential law. If the observations are sampled at equidistant times, then exponential decrease is equivalent to decrease by a constant fraction $0 < \Delta < 1$at each time step. Setting $w = 1 - \Delta$we can define $m$normalized weights by

$$w_i = \frac{w^{i-1}}{V_1},$$

where $V_1$is the sum of the unnormalized weights. In this case $V_1$is simply

$$V_1 = \sum_{i=1}^{m} w^{i-1} = \frac{1 - w^m}{1 - w},$$

approaching $V_1 = 1/(1 - w)$for large values of m.

The damping constant $w$must correspond to the actual decrease of interaction strength. If this cannot be determined from theoretical considerations, then the following properties of exponentially decreasing weights are useful in making a suitable choice: at step $(1-w)^{-1}$, the weight approximately equals $e^{-1}(1-w) = 0.39(1-w)$, the tail area the value e^{-1}, the head area $1 - e^{-1} = 0.61$. The tail area at step $n$is $\leq e^{-n(1-w)}$. Where primarily the closest $n$observations matter and the effect of the remaining observations can be ignored safely, then choose $w$such that the tail area is sufficiently small.

4.5.13 **Weighted averages of functions**

The concept of weighted average can be extended to functions. Weighted averages of functions play an important role in the systems of weighted differential and integral calculus.

4.6 Truncated mean

A **truncated mean** or **trimmed mean** is a statistical measure of central tendency, much like the mean and median. It involves the calculation of the mean after discarding given parts of a probability distribution or sample at the high and low end, and typically discarding an equal amount of both. This is usually given as a percentage, but may be given as a fixed number of points.

For most statistical applications, 5 to 25 percent of the ends are discarded; the 25% trimmed mean (when the lowest 25% and the highest 25% are discarded) is known as the interquartile mean. For example, given a set of 8 points, trimming by 12.5% would discard the minimum and maximum value in the sample: the first and last values.

The median can be regarded as a fully truncated mean and is most robust. As with other trimmed estimators, the main advantage of the trimmed mean is robustness and higher efficiency for mixed distributions and heavy-tailed distribution (like the Cauchy distribution), at the cost of lower efficiency for some other less heavily-tailed distributions (such as the normal distribution). For intermediate distributions the differences between the efficiency of the mean and the median are not very big, e.g. for the student-t distribution with 2 degrees of freedom the variances for mean and median are nearly equal.

4.6.1**Terminology**

In some regions of Central Europe it is also known as a **Windsor mean**, but this name should not be confused with the Winsorized mean: in the latter, the observations that the trimmed mean would discard are instead replaced by the largest/smallest of the remaining values.

Discarding only the maximum and minimum is known as the **modified mean**, particularly management statistics.

4.6.2 **Interpolation**

When the percentage of points to discard does not yield a whole number, the trimmed mean may be defined by interpolation, generally linear interpolation, between the nearest whole numbers. For example, if you need to calculate the 15% trimmed mean of a sample containing 10 entries, strictly this would mean discarding 1 point from each end (equivalent to the 10% trimmed mean). If interpolating, one would instead compute the 10% trimmed mean (discarding 1 point from each end) and the 20% trimmed mean

(discarding 2 points from each end), and then interpolating, in this case averaging these two values. Similarly, if interpolating the 12% trimmed mean, one would take the weighted average: weight the 10% trimmed mean by 0.8 and the 20% trimmed mean by 0.2.

4.6.3 **Advantages**

The truncated mean is a useful estimator because it is less sensitive to outliers than the mean but will still give a reasonable estimate of central tendency or mean for many statistical models. In this regard it is referred to as a robust estimator.

One situation in which it can be advantageous to use a truncated mean is when estimating the location parameter of a Cauchy distribution, a bell shaped probability distribution with (much) fatter tails than a normal distribution. It can be shown that the truncated mean of the middle 24% sample order statistics (i.e., truncate the sample by 38%) produces an estimate for the population location parameter that is more efficient than using either the sample median or the full sample mean. However, due to the fat tails of the Cauchy distribution, the efficiency of the estimator decreases as more of the sample gets used in the estimate. Note that for the Cauchy distribution, neither the truncated mean, full sample mean or sample median represents a maximum likelihood estimator, nor are any as asymptotically efficient as the maximum likelihood estimator; however, the maximum likelihood estimate is more difficult to compute, leaving the truncated mean as a useful alternative.

4.6.4 **Drawbacks**

The truncated mean uses more information from the distribution or sample than the median, but unless the underlying distribution is symmetric, the truncated mean of a sample is unlikely to produce an unbiased estimator for either the mean or the median.

4.6.5 **Examples**

The scoring method used in many sports that are evaluated by a panel of judges is a truncated mean: *discard the lowest and the highest scores; calculate the mean value of the remaining scores*.

The Libor benchmark interest rate is calculated as a trimmed mean: given 18 response, the top 4 and bottom 4 are discarded, and the remaining 10 are averaged (yielding trim factor of $4/18 \approx 22\%$).

4.7 **Interquartile mean**

The **interquartile mean (IQM)** (or **midmean**) is a statistical measure of central tendency, much like the mean (in more popular terms called the average), the median, and the mode.

The IQM is a *truncated mean* and so is very similar to the scoring method used in sports that are evaluated by a panel of judges: *discard the lowest and the highest scores; calculate the mean value of the remaining scores*.

4.7.1**Calculation**

In calculation of the IQM, only the data in the second and third quantiles is used (as in the interquartile range), and the lowest 25% and the highest 25% of the scores are discarded. These points are called the

first and third quartiles, hence the name of the IQM. (Note that the *second* quartile is also called the median).

$$x_{\mathrm{IQM}} = \frac{2}{n} \sum_{i=\frac{n}{4}+1}^{\frac{3n}{4}} x_i$$

assuming the values have been ordered.

4.7.2 **Dataset divisible by four**

The method is best explained with an example. Consider the following dataset:

5, 8, 4, 38, 8, 6, 9, 7, 7, 3, 1, 6

First sort the list from lowest-to-highest:

1, 3, 4, 5, 6, 6, 7, 7, 8, 8, 9, 38

There are 12 observations (datapoints) in the dataset, thus we have 4 quartiles of 3 numbers. Discard the lowest and the highest 3 values:

~~1, 3, 4~~, 5, 6, 6, 7, 7, 8, ~~8, 9, 38~~

We now have 6 of the 12 observations remaining; next, we calculate the arithmetic mean of these numbers:

x_{IQM} = (5 + 6 + 6 + 7 + 7 + 8) / 6 = 6.5

For comparison, the arithmetic mean of the original dataset is

(5 + 8 + 4 + 38 + 8 + 6 + 9 + 7 + 7 + 3 + 1 + 6) / 12 = 8.5

due to the strong influence of the outlier, 38.

4.7.3 **Dataset not divisible by four**

The above example consisted of 12 observations in the dataset, which made the determination of the quartiles very easy. Of course, not all datasets have a number of observations that is divisible by 4. We can adjust the method of calculating the IQM to accommodate this. So ideally we want to have the IQM equal to the mean for symmetric distributions, e.g.:

1, 2, 3, 4, 5

has a mean value x_{mean} = 3, and since it is a symmetric distribution, x_{IQM} = 3 would be desired.

We can solve this by using a weighted average of the quartiles and the interquartile dataset:

Consider the following dataset of 9 observations:

1, 3, 5, 7, 9, 11, 13, 15, 17

There are 9/4 = 2.25 observations in each quartile, and 4.5 observations in the interquartile range. Truncate the fractional quartile size, and remove this number from the 1st and 3rd quartiles (2.25 observations in each quartile, thus the lowest 2 and the highest 2 are removed).

~~1, 3~~, (5), 7, 9, 11, (13), ~~15, 17~~

Thus, there are 3 *full* observations in the interquartile range, and 2 fractional observations. Since we have a total of 4.5 observations in the interquartile range, the two fractional observations each count for 0.75 (and thus 3×1 + 2×0.75 = 4.5 observations).

The IQM is now calculated as follows:

$x_{\mathrm{IQM}} = \{(7 + 9 + 11) + 0.75 \times (5 + 13)\} / 4.5 = 9$

In the above example, the mean has a value $x_{\mathrm{mean}} = 9$. The same as the IQM, as was expected. The method of calculating the IQM for any number of observations is analogous; the fractional contributions to the IQM can be either 0, 0.25, 0.50, or 0.75.

4.7.4 **Comparison with mean and median**

The Interquartile Mean shares some properties from both the mean as well as the median:

- Like the median, the IQM is insensitive to outliers; in the example given, the highest value (38) was an obvious outlier of the dataset, but its value is not used in the calculation of the IQM. On the other hand, the common average (the arithmetic mean) is sensitive to these outliers: $x_{\mathrm{mean}} = 8.5$.
- Like the mean, the IQM is a discrete parameter, based on a large number of observations from the dataset. The median is always equal to *one* of the observations in the dataset (assuming an odd number of observations). The mean can be equal to *any* value between the lowest and highest observation, depending on the value of *all* the other observations. The IQM can be equal to *any* value between the first and third quartiles, depending on *all* the observations in the interquartile range.

4.8 **Mid-range**

In statistics, the **mid-range** or **mid-extreme** of a set of statistical data values is the arithmetic mean of the maximum and minimum values in a data set, defined as:

$$M = \frac{\max x + \min x}{2}.$$

The mid-range is the midpoint of the range; as such, it is a measure of central tendency.

The mid-range is rarely used in practical statistical analysis, as it lacks efficiency as an estimator for most distributions of interest, because it ignores all intermediate points, and lacks robustness, as outliers change

it significantly. Indeed, it is one of the least efficient and least robust statistics. However, it finds some use in special cases: it is the maximally efficient estimator for the center of a uniform distribution, trimmed mid-ranges address robustness, and as an L-estimator, it is simple to understand and compute.

4.8.1 **Comparison with other measures**

4.8.1.1 **Robustness**

The midrange is highly sensitive to outliers and ignores all but two data points. It is therefore a very non-robust statistic, having a breakdown point of 0, meaning that a single observation can change it arbitrarily. Further, it is highly influenced by outliers: increasing the sample maximum or decreasing the sample minimum by x changes the mid-range by $x/2$, while it changes the sample mean, which also has breakdown point of 0, by only x/n. It is thus of little use in practical statistics, unless outliers are already handled.

A trimmed midrange is known as a **midsummary** – the n% trimmed midrange is the average of the n% and $(100-n)$% percentiles, and is more robust, having a breakdown point of n%. In the middle of these is the midhinge, which is the 25% midsummary. The median can be interpreted as the fully trimmed (50%) mid-range; this accords with the convention that the median of an even number of points is the mean of the two middle points.

These trimmed midranges are also of interest as descriptive statistics or as L-estimators of central location or skewness: differences of midsummaries, such as midhinge minus the median, give measures of skewness at different points in the tail.

4.8.1.2 **Efficiency**

Despite its drawbacks, in some cases it is useful: the midrange is a highly efficient estimator of μ, given a small sample of a sufficiently platykurtic distribution, but it is inefficient for mesokurtic distributions, such as the normal.

For example, for a continuous uniform distribution with unknown maximum and minimum, the mid-range is the UMVU estimator for the mean. The sample maximum and sample minimum, together with sample size, are a sufficient statistic for the population maximum and minimum – the distribution of other samples, conditional on a given maximum and minimum, is just the uniform distribution between the maximum and minimum and thus add no information. Thus the mid-range, which is an unbiased and sufficient estimator of the population mean, is in fact the UMVU: using the sample mean just adds noise based on the uninformative distribution of points within this range.

Conversely, for the normal distribution, the sample mean is the UMVU estimator of the mean. Thus for platykurtic distributions, which can often be thought of as between a uniform distribution and a normal distribution, the informativeness of the middle sample points versus the extrema values varies from "equal" for normal to "uninformative" for uniform, and for different distributions, one or the other (or some combination thereof) may be most efficient. A robust analog is the trimean, which averages the midhinge (25% trimmed mid-range) and median.

4.8.1.3 ***Small samples***

For small sample sizes (*n* from 4 to 20) drawn from a sufficiently platykurtic distribution (negative excess kurtosis, defined as $\gamma_2 = (\mu_4/(\mu_2)^2) - 3$), the mid-range is an efficient estimator of the mean μ. The following table summarizes empirical data comparing three estimators of the mean for distributions of varied kurtosis; the modified mean is the truncated mean, where the maximum and minimum are eliminated.

Excess kurtosis (γ_2)	**Most efficient estimator of μ**
−1.2 to −0.8	Midrange
−0.8 to 2.0	Mean
2.0 to 6.0	Modified mean

For $n = 1$ or 2, the midrange and the mean are equal (and coincide with the median), and are most efficient for all distributions. For $n = 3$, the modified mean is the median, and instead the mean is the most efficient measure of central tendency for values of γ_2 from 2.0 to 6.0 as well as from −0.8 to 2.0.

4.8.2 **Sampling properties**

For a sample of size *n* from the standard normal distribution, the mid-range *M* is unbiased, and has a variance given by:

$$\operatorname{var}(M) = \frac{\pi^2}{24\ln(n)}.$$

For a sample of size *n* from the standard Laplace distribution, the mid-range *M* is unbiased, and has a variance given by:

$$\operatorname{var}(M) = \frac{\pi^2}{12}$$

and, in particular, the variance does not decrease to zero as the sample size grows.

For a sample of size *n* from a zero-centred uniform distribution, the mid-range *M* is unbiased, *nM* has an asymptotic distribution which is a Laplace distribution.

4.8.3 **Deviation**

While the mean of a set of values minimizes the sum of squares of deviations and the median minimizes the average absolute deviation, the midrange minimizes the maximum deviation (defined as $\max|x_i - m|$): it is a solution to a variational problem.

4.9 Absolute deviation

In statistics, the **absolute deviation** of an element of a data set is the absolute difference between that element and a given point. Typically the deviation is reckoned from the central value, being construed as some type of average, most often the median or sometimes the mean of the data set.

$$D_i = |x_i - m(X)|$$

where

D_i is the absolute deviation,
x_i is the data element
and $m(X)$ is the chosen measure of central tendency of the data set—sometimes the mean ($\overline{x}$), but most often the median.

4.9.1 **Measures of dispersion**

Several measures of statistical dispersion are defined in terms of the absolute deviation.

4.9.2 **Average absolute deviation**

The **average absolute deviation,** or simply **average deviation** of a data set is the average of the absolute deviations and is a summary statistic of statistical dispersion or variability. In its general form, the average used can be the mean, median, mode, or the result of another measure of central tendency.

The average absolute deviation of a set $\{x_1, x_2, ..., x_n\}$ is

$$\frac{1}{n}\sum_{i=1}^{n}|x_i - m(X)|.$$

The choice of measure of central tendency, $m(X)$, has a marked effect on the value of the average deviation. For example, for the data set {2, 2, 3, 4, 14}:

Measure of central tendency $m(X)$	**Average absolute deviation**
Mean = 5	$\frac{\|2-5\| + \|2-5\| + \|3-5\| + \|4-5\| + \|14-5\|}{5} = 3.6$
Median = 3	$\frac{\|2-3\| + \|2-3\| + \|3-3\| + \|4-3\| + \|14-3\|}{5} = 2.8$
Mode = 2	$\frac{\|2-2\| + \|2-2\| + \|3-2\| + \|4-2\| + \|14-2\|}{5} = 3.0$

The average absolute deviation from the median is less than or equal to the average absolute deviation from the mean. In fact, the average absolute deviation from the median is always less than or equal to the average absolute deviation from any other fixed number.

The average absolute deviation from the mean is less than or equal to the standard deviation; one way of proving this relies on Jensen's inequality.

For the normal distribution, the ratio of mean absolute deviation to standard deviation is $\sqrt{2/\pi}=0.79788456...$. Thus if *X* is a normally distributed random variable with expected value 0 then

$$\frac{E|X|}{\sqrt{E(X^2)}} = \sqrt{\frac{2}{\pi}}.$$

In other words, for a normal distribution, mean absolute deviation is about 0.8 times the standard deviation.

4.9.3 ***Mean absolute deviation (MAD)***

The **mean absolute deviation** (MAD), also referred to as the mean deviation, is the mean of the absolute deviations of a set of data about the data's mean. In other words, it is the average distance of the data set from its mean. Because the MAD is a simpler measure of variability than the standard deviation, it can be used as pedagogical tool to help motivate the standard deviation.

This method forecast accuracy is very closely related to the mean squared error (MSE) method which is just the average squared error of the forecasts. Although these methods are very closely related MAD is more commonly used because it does not require squaring.

More recently, the mean absolute deviation about mean is expressed as a covariance between a random variable and its under/over indicator functions; as

$$D_m = E|X - \mu| = 2Cov(X, I_O)$$

where

D_m is the expected value of the absolute deviation about mean,
"Cov" is the covariance between the random variable X and the over indicator function (I_O).

and the over indicator function is defined as

$$\mathbf{I}_O := \begin{cases} 1 & \text{if } x > \mu, \\ 0 & \text{else} \end{cases}$$

Based on this representation new correlation coefficients are derived. These correlation coefficients ensure high stability of statistical inference when we deal with distributions that are not symmetric and for which the normal distribution is not an appropriate approximation. Moreover an easy and simple way for a semi decomposition of Pietra's index of inequality is obtained.

4.9.4 **Average absolute deviation about median**

Mean absolute deviation about median (MAD median) offers a direct measure of the scale of a random variable about its median

$$D_{med} = E|X - median|$$

For the normal distribution we have $D_{med} = \sigma\sqrt{(2/\pi)}$. Since the median minimizes the average absolute distance, we have $D_{med} <= D_{mean}$. By using the general dispersion function Habib (2011) defined MAD about median as

$$D_{med} = E|X - median| = 2Cov(X, I_O)$$

where the indicator function is

$$\mathbf{I}_O := \begin{cases} 1 & \text{if } x > median, \\ 0 & \text{else} \end{cases}$$

This representation allows for obtaining MAD median correlation coefficients;

4.9.5 ***Median absolute deviation (MAD)***

The **median absolute deviation** (also MAD) is the *median* of the absolute deviation from the *median*. It is a robust estimator of dispersion.

For the example {2, 2, 3, 4, 14}: 3 is the median, so the absolute deviations from the median are {1, 1, 0, 1, 11} (reordered as {0, 1, 1, 1, 11}) with a median of 1, in this case unaffected by the value of the outlier 14, so the median absolute deviation (also called MAD) is 1.

4.9.6 ***Maximum absolute deviation***

The **maximum absolute deviation** about a point is the maximum of the absolute deviations of a sample from that point. While not strictly a measure of central tendency, the maximum absolute deviation can be found using the formula for the average absolute deviation as above with $m(X) = \max(X)$, where $\max(X)$is the sample maximum. The maximum absolute deviation cannot be less than half the range.

4.9.7 **Minimization**

The measures of statistical dispersion derived from absolute deviation characterize various measures of central tendency as *minimizing* dispersion: The median is the measure of central tendency most associated with the absolute deviation. Some location parameters can be compared as follows:

- L^2 norm statistics: the mean minimizes the mean squared error
- L^1 norm statistics: the median minimizes *average* absolute deviation,
- L^∞ norm statistics: the mid-range minimizes the *maximum* absolute deviation

- trimmed L^∞ norm statistics: for example, the midhinge (average of first and third quartiles) which minimizes the *median* absolute deviation of the whole distribution, also minimizes the *maximum* absolute deviation of the distribution after the top and bottom 25% have been trimmed off.

4.9.8 Estimation

The mean absolute deviation of a sample is a biased estimator of the mean absolute deviation of the population. In order for the absolute deviation to be an unbiased estimator, the expected value (average) of all the sample absolute deviations must equal the population absolute deviation. However, it does not. For the population 1,2,3 both the population absolute deviation about the median and the population absolute deviation about the mean are 2/3. The average of all the sample absolute deviations about the mean of size 3 that can be drawn from the population is 44/81, while the average of all the sample absolute deviations about the median is 4/9. Therefore the absolute deviation is a biased estimator.

4.10 Nonparametric skew

In statistics and probability theory, the **nonparametric skew** is a statistic occasionally used with random variables that take real values. It is a measure of the skewness of a random variable's distribution—that is, the distribution's tendency to "lean" to one side or the other of the mean. Its calculation does not require any knowledge of the form of the underlying distribution—hence the name nonparametric. It has some desirable properties: it is zero for any symmetric distribution; it is unaffected by a scale shift; and it reveals either left- or right-skewness equally well. Although its use has been mentioned in older textbooks it appears to have gone out of fashion. In statistical samples it has been shown to be less powerful than the usual measures of skewness in detecting departures of the population from normality.

4.10.1 **Definition**

The nonparametric skew is defined as

$$S = \frac{\mu - \nu}{\sigma}$$

where the mean (μ), median (ν) and standard deviation (σ) of the population have their usual meanings.

4.10.2 **Properties**

The nonparametric skew is one third of the Pearson 2 skewness coefficient and lies between −1 and +1 for any distribution. This range is implied by the fact that the mean lies within one standard deviation of any median.

Under an affine transformation of the variable (X), the value of S does not change except for a possible change in sign. In symbols

$$S(aX + b) = \operatorname{sign}(a)\, S(X)$$

where $a \neq 0$ and b are constants and $S(X)$ is the nonparametric skew of the variable X.

4.10.3 **Sharper bounds**

The bounds of this statistic (±1) were sharpened by Majindar who showed that its absolute value is bounded by

$$\frac{2(pq)^{1/2}}{(p+q)^{1/2}}$$

with

$$p = \Pr(X > \mathrm{E}(X))$$

and

$$q = \Pr(X < \mathrm{E}(X)),$$

where X is a random variable with finite variance, $E()$ is the expectation operator and $Pr()$ is the probability of the event occurring.

When $p = q = 0.5$ the absolute value of this statistic is bounded by 1. With $p = 0.1$ and $p = 0.01$, the statistic's absolute value is bounded by 0.6 and 0.199 respectively.

4.10.4 **Extensions**

It is also known that

$$|\mu - \nu_0| \leq \mathrm{E}(|X - \nu_0|) \leq \mathrm{E}(|X - \mu|) \leq \sigma,$$

where v_0 is any median and $E(.)$ is the expectation operator.

It has been shown that

$$\frac{|\mu - x_q|}{\sigma} \leq \max\left(\sqrt{\frac{(1-q)}{q}}, \sqrt{\frac{q}{(1-q)}}\right)$$

where x_q is the q^{th} quantile. Quantiles lie between 0 and 1: the median (the 0.5 quantile) has $q = 0.5$. This inequality has also been used to define a measure of skewness.

This latter inequality has been sharpened further.

$$\mu - \sigma\sqrt{\frac{1-q}{q}} \leq x_q \leq \mu + \sigma\sqrt{\frac{q}{1-q}}$$

Another extension for a distribution with a finite mean has been published:

$$\mu - \frac{1}{2q}\,\mathrm{E}\,|X-\mu| \le x_q \le \mu + \frac{1}{(2-2q)}\,\mathrm{E}\,|X-\mu|$$

The bounds in this last pair of inequalities are attained when $\Pr(X=a)=q$ and $\Pr(X=b)=1-q$ for fixed numbers $a < b$.

4.10.5 **Finite samples**

For a finite sample with sample size $n \ge 2$ with x_r is the r^{th} order statistic, m the sample mean and s the sample standard deviation corrected for degrees of freedom,

$$\frac{|m-x_r|}{s} \le \max\left[\sqrt{\frac{(n-1)(r-1)}{n(n-r+1)}}, \sqrt{\frac{(n-1)(n-r)}{nr}}\right]$$

Replacing r with $n / 2$ gives the result appropriate for the sample median:

$$\frac{|m-a|}{s} \le \sqrt{\frac{n^2-n}{n^2}} = \sqrt{\frac{n-1}{n}}$$

where a is the sample median.

4.10.6 **Statistical tests**

Hotelling and Solomons considered the distribution of the test statistic

$$D = \frac{n(m-a)}{s}$$

where n is the sample size, m is the sample mean, a is the sample median and s is the sample's standard deviation.

Statistical tests of D have assumed that the null hypothesis being tested is that the distribution is symmetric .

Gastwirth estimated the asymptotic variance of $n^{-1/2}D$. If the distribution is unimodal and symmetric about 0, the asymptotic variance lies between 1/4 and 1. Assuming a conservative estimate (putting the variance equal to 1) can lead to a true level of significance well below the nominal level.

Assuming that the underlying distribution is symmetric Cabilio and Masaro have shown that the distribution of S is asymptotically normal. The asymptotic variance depends on the underlying distribution: for the normal distribution, the asymptotic variance of ($S\sqrt{n}$) is 0.5708.

Assuming that the underlying distribution is symmetric, by considering the distribution of values above and below the median Zheng and Gastwirth have argued that

$$\sqrt{2n}(\frac{m-a}{s})$$

where *n* is the sample size, is distributed as a t distribution.

4.10.7 **Related statistics**

Mira studied the distribution of the difference between the mean and the median.

$$\gamma_1 = 2(m-a),$$

where *m* is the sample mean and *a* is the median. If the underlying distribution is symmetrical γ_1 itself is asymptotically normal. This statistic had been earlier suggested by Bonferroni.

Assuming a symmetric underlying distribution, a modification of *S* was studied by Miao, Gel and Gastwirth who modified the standard deviation to create their statistic.

$$J = \frac{1}{n}\sqrt{\frac{\pi}{2}}\sum |X_i - a|$$

where X_i are the sample values, || is the absolute value and the sum is taken over all *n* sample values.

The test statistic was

$$T = \frac{m-a}{J}.$$

The scaled statistic ($T\sqrt{n}$) is asymptotically normal with a mean of zero for a symmetric distribution. Its asymptotic variance depends on the underlying distribution: the limiting values are, for the normal distribution var($T\sqrt{n}$) = 0.5708 and, for the t distribution with three degrees of freedom, var($T\sqrt{n}$) = 0.9689.

4.10.8 **Values for individual distributions**

4.10.8.1 **Symmetric distributions**

For symmetric probability distributions the value of the nonparametric skew is 0.

4.10.8.2 **Asymmetric distributions**

It is positive for right skewed distributions and negative for left skewed distributions. Absolute values ≥ 0.2 indicate marked skewness.

It may be difficult to determine *S* for some distributions. This is usually because a closed form for the median is not known: examples of such distributions include the gamma distribution, inverse-chi-squared distribution, the inverse-gamma distribution and the scaled inverse chi-squared distribution.

The following values for *S* are known:

- Beta distribution: $1 < \alpha < \beta$ where α and β are the parameters of the distribution, then to a good approximation

  $$S = \frac{1}{3}\frac{(\alpha - 2\beta)(\alpha + \beta + 1)^{1/2}}{(\alpha + \beta - 2/3)(\alpha\beta)^{1/2}}$$

  If $1 < \beta < \alpha$ then the positions of α and β are reversed in the formula. *S* is always < 0.

- Binomial distribution: varies. If the mean is an integer then $S = 0$. If the mean is not an integer *S* may have either sign or be zero. It is bounded by $\pm\min\{ \max\{ p, 1 - p \}, \log_e 2 \} / \sigma$ where σ is the standard deviation of the binomial distribution.

- Burr distribution:

- Birnbaum–Saunders distribution:

  $$S = \frac{2}{\beta^2(4 + 5\alpha^2)}$$

  where α is the shape parameter and β is the location parameter.

- Cantor distribution:

  $$\frac{-4}{3} \leq S \leq \frac{4}{3}$$

- Chi square distribution: Although $S \geq 0$ its value depends on the numbers of degrees of freedom (k).

  $$S \approx \frac{1 - (1 - \frac{2}{k})^3}{2}$$

- Dagum distribution:

- Exponential distribution:

  $$S = 1 - \log_e(2) \approx 0.31$$

- Exponential distribution with two parameters:

  $$S = 1 - \log_e(2) \approx 0.31$$

- Exponential-logarithmic distribution

$$S = -\frac{polylog(2, 1-p) + \ln(1+\sqrt{p})\ln p}{\sqrt{-[2polylog(3, 1-p) + polylog^2(2, 1-p)]}}$$

Here S is always > 0.

- Exponentially modified Gaussian distribution:

$$0 \leq S \leq 1 - \log_e(2)$$

- F distribution with n and n degrees of freedom ($n > 4$):

$$S = n^{-3/2}\sqrt{\frac{n-4}{n-2}} + O(n^{-5/2})$$

- Fréchet distribution: The variance of this distribution is defined only for $\alpha > 2$.

$$S = \frac{\Gamma\left(1-\frac{1}{\alpha}\right) - \frac{1}{\sqrt{\alpha \log_e(2)}}}{\sqrt{\Gamma\left(1-\frac{2}{\alpha}\right) - \left(\Gamma\left(1-\frac{1}{\alpha}\right)\right)^2}}$$

- Gamma distribution: The median can only be determined approximately for this distribution. If the shape parameter α is ≥ 1 then

$$S \approx \frac{\beta}{3\alpha + 0.2}$$

where $\beta > 0$ is the rate parameter. Here S is always > 0.

- Generalized normal distribution version 2

$$S = -\frac{\exp(\frac{-k^2}{2}) - 1}{\sqrt{\exp(\frac{k^2}{2}) - 1}}$$

S is always < 0.

- Generalized Pareto distribution: S is defined only when the shape parameter (k) is < 1/2. S is < 0 for this distribution.

$$S = \left(\frac{2^k - 1}{k} - 2^k\right)(1-2k)^{0.5}$$

- Gumbel distribution:

$$\frac{\sqrt{6}[\gamma + \log_e(\log_e(2))]}{\pi} \approx 0.1643$$

where γ is Euler's constant.

- Half-normal distribution:

$$S \approx \frac{\sqrt{2} - 0.6745\sqrt{\pi}}{\sqrt{\pi - 2}} \approx 0.36279$$

- Kumaraswamy distribution

- Log-logistic distribution (Fisk distribution): Let β be the shape parameter. The variance and mean of this distribution are only defined when $\beta > 2$. To simplify the notation let $b = \beta / \pi$.

$$S = \frac{b - \sin(b)}{\sqrt{b\tan(b) - b^2}}$$

The standard deviation does not exist for values of $b > 4.932$ (approximately). For values for which the standard deviation is defined, S is > 0.

- Log-normal distribution: With mean (μ) and variance (σ^2)

$$S = \frac{1}{(e^{\frac{\sigma^2}{2}} + 1)(e^{\mu+\sigma^2})}$$

- Log-Weibull distribution:

$$S \approx \frac{[\log_e(\log_e(2)) - 0.5772]\sqrt{6}}{\pi} \approx -0.1643$$

- Lomax distribution: S is defined only for $\alpha > 2$

$$S = \frac{(\alpha - 1)(\alpha - 2)(1 - (\alpha - 1)(2^{1/\alpha} - 1))}{\alpha^{1/2}}$$

- Maxwell–Boltzmann distribution:

$$S \approx \frac{\sqrt{2} - 1.5382\Gamma(\frac{3}{2})}{\sqrt{2(\Gamma(\frac{5}{2}) - \Gamma(\frac{3}{2}))}} \approx 0.0854$$

- Nakagami distribution

$$S = -1$$

- Pareto distribution: for $\alpha > 2$ where α is the shape parameter of the distribution,

$$S = (\alpha - 2^{1/\alpha}[\alpha - 1])(\frac{\alpha - 2}{\alpha})^{1/2},$$

and S is always > 0.

- Poisson distribution:

$$\frac{-\log_e(2)}{\lambda^{\frac{1}{2}}} \leq S \leq \frac{1}{3\lambda^{\frac{1}{2}}}$$

where λ is the parameter of the distribution.

- Rayleigh distribution:

$$S = \sqrt{\frac{2}{4-\pi}}[(\frac{\pi}{2})^{0.5} - \log_e(4)] \approx 0.1251$$

- Weibull distribution:

$$S = \frac{\Gamma(1+1/k) - \log_e(2)^{1/k}}{(\Gamma(1+2/k) - \Gamma(1+1/k))^{1/2}},$$

where k is the shape parameter of the distribution. Here S is always > 0.

4.10.9 **History**

In 1895 Pearson first suggested suggested measuring skewness by standardizing the difference between the mean and the mode, giving

$$\frac{\mu - \theta}{\sigma},$$

where μ, θ and σ is the mean, mode and standard deviation of the distribution respectively. Estimates of the population mode from the sample data may be difficult but the difference between the mean and the mode for many distributions is approximately three times the difference between the mean and the median which suggested to Pearson a second skewness coefficient:

$$\frac{3(\mu - \nu)}{\sigma},$$

where v is the median of the distribution. Bowley dropped the factor 3 is from this formula in 1901 leading to the nonparametric skew statistic.

The relationship between the median, the mean and the mode was first noted by Pearson when he was investigating his type III distributions.

4.10.10 **Relationships between the mean, median and mode**

For an arbitrary distribution the mode, median and mean may appear in any order.

Analyses have been made of some of the relationships between the mean, median, mode and standard deviation. and these relationships place some restrictions of the sign and magnitude of the nonparametric skew.

A simple example illustrating these relationships is the binomial distribution with $n = 10$ and $p = 0.09$. This distribution when plotted has a long right tail. The mean (0.9) is to the left of the median (1) but the skew (0.906) as defined by the third standardized moment is positive. In contrast the nonparametric skew is -0.110.

4.10.11 **Pearson's rule**

The rule that for some distributions the difference between the mean and the mode is three times that between the mean and the median is due to Pearson who discovered it while investigating his Type 3 distributions. It is often applied to slightly asymmetric distributions that resemble a normal distribution but it is not always true.

In 1895 Pearson noted that for what is now known as the gamma distribution that the relation

$$\nu - \theta = 2(\mu - \nu)$$

where θ, ν and μ are the mode, median and mean of the distribution respectively was approximately true for distributions with a large shape parameter.

Doodson in 1917 proved that the median lies between the mode and the mean for moderately skewed distributions with finite fourth moments. This relationship holds for all the Pearson distributions and all of these distributions have a positive nonparametric skew.

Doodson also noted that for this family of distributions to a good approximation,

$$\theta = 3\nu - 2\mu,$$

where θ, ν and μ are the mode, median and mean of the distribution respectively. Doodson's approximation was further investigated and confirmed by Haldane. Haldane noted that in samples with identical and independent variates with a third cumulant had sample means that obeyed Pearson's relationship for large sample sizes. Haldane required a number of conditions for this relationship to hold including the existence of an Edgeworth expansion and the uniqueness of both the median and the mode. Under these conditions he found that mode and the median converged to 1/2 and 1/6 of the third moment respectively. This result was confirmed by Hall under weaker conditions using characteristic functions.

Doodson's relationship was studied by Kendall and Stuart in the log-normal distribution for which they found an exact relationship close to it.

Hall also showed that for a distribution with regularly varying tails and exponent α that

$$\mu - \theta = \alpha(\mu - \nu)$$

4.10.12 **Unimodal distributions**

Gauss showed in 1823 that for a unimodal distribution

$$\sigma \leq \omega \leq 2\sigma$$

and

$$|\nu - \mu| \leq \sqrt{\frac{3}{4}}\omega,$$

where ω is the root mean square deviation from the mode.

For a large class of unimodal distributions that are positively skewed the mode, median and mean fall in that order. Conversely for a large class of unimodal distributions that are negatively skewed the mean is less than the median which in turn is less than the mode. In symbols for these positively skewed unimodal distributions

$$\theta \leq \nu \leq \mu$$

and for these negatively skewed unimodal distributions

$$\mu \leq \nu \leq \theta$$

This class includes the important F, beta and gamma distributions.

This rule does not hold for the unimodal Weibull distribution.

For a unimodal distribution the following bounds are known and are sharp:

$$\frac{|\theta - \mu|}{\sigma} \leq \sqrt{3},$$
$$\frac{|\nu - \mu|}{\sigma} \leq \sqrt{0.6},$$
$$\frac{|\theta - \nu|}{\sigma} \leq \sqrt{3},$$

where μ,ν and θ are the mean, median and mode respectively.

The middle bound limits the nonparametric skew of a unimodal distribution to approximately ±0.775.

4.10.13 **van Zwet condition**

The following inequality,

$$\theta \leq \nu \leq \mu,$$

where θ, ν and μ is the mode, median and mean of the distribution respectively, holds if

$$F(\nu - x) + F(\nu + x) \geq 1 \text{ for all } x,$$

where F is the cumulative distribution function of the distribution. These conditions have since been generalised and extended to discrete distributions. Any distribution for which this holds has either a zero or a positive nonparametric skew.

4.10.14 **Notes**

4.10.14.1 **Ordering of skewness**

In 1964 van Zwet proposed a series of axioms for ordering measures of skewness. The nonparametric skew does not satisfy these axioms.

4.10.14.2 **Benford's law**

Benford's law is an empirical law concerning the distribution of digits in a list of numbers. It has been suggested that random variates from distributions with a positive nonparametric skew will obey this law.

4.10.14.3 **Relation to Bowley's coefficient**

This statistic can be derived from Bowley's coefficient of skewness

$$SK_2 = \frac{Q_3 + Q_1 - 2Q_2}{Q_3 - Q_1}$$

where Q_i is the ith quartile of the distribution.

Hinkley generalised this

$$SK = \frac{F(1-\alpha) + F(\alpha) - 2Q_2}{Q_3 - Q_1}$$

where lies between 0 and 0.5. Bowley's coefficient is a special case with equal to 0.25.

Groeneveld and Meeden removed the dependence on by integrating over it.

$$SK_3 = \frac{\mu - 2Q_2}{E|y - Q_2|}$$

The denominator is a measure of dispersion. Replacing the denominator with the standard deviation we obtain the nonparametric skew.

4.11 Median

In statistics and probability theory, the **median** is the numerical value separating the higher half of a data sample, a population, or a probability distribution, from the lower half. The *median* of a finite list of numbers can be found by arranging all the observations from lowest value to highest value and picking the middle one (e.g., the median of {3, 5, 9} is 5). If there is an even number of observations, then there is no single middle value; the median is then usually defined to be the mean of the two middle values, which corresponds to interpreting the median as the fully trimmed mid-range. The median is of central importance in robust statistics, as it is the most resistant statistic, having a breakdown point of 50%: so long as no more than half the data is contaminated, the median will not give an arbitrarily large result.

A median is only defined on ordered one-dimensional data, and is independent of any distance metric. A geometric median, on the other hand, is defined in any number of dimensions.Also a median can be regarded the "middle number" in a set of values and in the case of having even number of values such as (3,4,5,6) the median will be the two middle values added together and divided by 2, [i.e (4+5)/2]

In a sample of data, or a finite population, there may be no member of the sample whose value is identical to the median (in the case of an even sample size); if there is such a member, there may be more than one so that the median may not uniquely identify a sample member. Nonetheless, the value of the median is uniquely determined with the usual definition. A related concept, in which the outcome is forced to correspond to a member of the sample, is the medoid. At most, half the population have values strictly less than the *median*, and, at most, half have values strictly greater than the median. If each group contains less than half the population, then some of the population is exactly equal to the median. For example, if $a < b < c$, then the median of the list $\{a, b, c\}$ is b, and, if $a < b < c < d$, then the median of the list $\{a, b, c, d\}$ is the mean of b and c; i.e., it is $(b + c)/2$.

The median can be used as a measure of location when a distribution is skewed, when end-values are not known, or when one requires reduced importance to be attached to outliers, e.g., because they may be measurement errors.

In terms of notation, some authors represent the median of a variable x either as $\tilde{x}$or as $\mu_{1/2}$, sometimes also M. There is no widely accepted standard notation for the median, so the use of these or other symbols for the median needs to be explicitly defined when they are introduced.

The median is the 2nd quartile, 5th decile, and 50th percentile.

4.11.1 **Measures of location and dispersion**

The median is one of a number of ways of summarising the typical values associated with members of a statistical population; thus, it is a possible location parameter.

When the median is used as a location parameter in descriptive statistics, there are several choices for a measure of variability: the range, the interquartile range, the mean absolute deviation, and the median absolute deviation. Since the median is the same as the *second quartile*, its calculation is illustrated in the article on quartiles.

For practical purposes, different measures of location and dispersion are often compared on the basis of how well the corresponding population values can be estimated from a sample of data. The median, estimated using the sample median, has good properties in this regard. While it is not usually optimal if a given population distribution is assumed, its properties are always reasonably good. For example, a comparison of the efficiency of candidate estimators shows that the sample mean is more statistically efficient than the sample median when data are uncontaminated by data from heavy-tailed distributions or

from mixtures of distributions, but less efficient otherwise, and that the efficiency of the sample median is higher than that for a wide range of distributions. More specifically, the median has a 64% efficiency compared to the minimum-variance mean (for large normal samples), which is to say the variance of the median will be ~50% greater than the variance of the mean.

4.11.2 **Medians of probability distributions**

For any probability distribution on the real line **R** with cumulative distribution function *F*, regardless of whether it is any kind of continuous probability distribution, in particular an absolutely continuous distribution (which has a probability density function), or a discrete probability distribution, a median is by definition any real number *m* that satisfies the inequalities

$$\mathrm{P}(X \leq m) \geq \frac{1}{2} \text{ and } \mathrm{P}(X \geq m) \geq \frac{1}{2}$$

or, equivalently, the inequalities

$$\int_{(-\infty, m]} dF(x) \geq \frac{1}{2} \text{ and } \int_{[m, \infty)} dF(x) \geq \frac{1}{2}$$

in which a Lebesgue–Stieltjes integral is used. For an absolutely continuous probability distribution with probability density function *f*, the median satisfies

$$\mathrm{P}(X \leq m) = \mathrm{P}(X \geq m) = \int_{-\infty}^{m} f(x)\,dx = \frac{1}{2}.$$

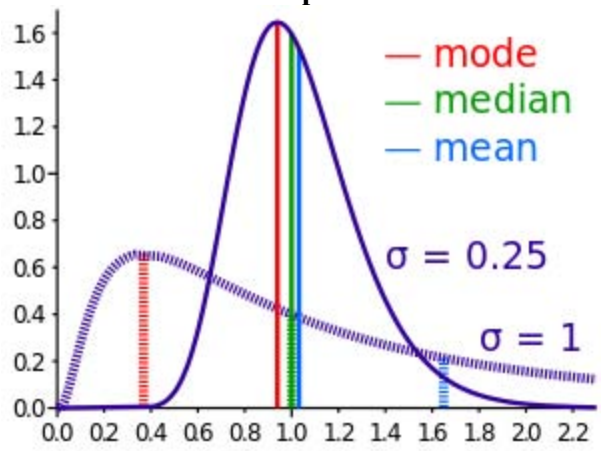
Any probability distribution on **R** has at least one median, but there may be more than one median. Where exactly one median exists, statisticians speak of "the median" correctly; even when the median is not unique, some statisticians speak of "the median" informally.

4.11.3 **Medians of particular distributions**

The medians of certain types of distributions can be easily calculated from their parameters:

- The median of a symmetric distribution with mean μ is μ.
  - The median of a normal distribution with mean μ and variance σ^2 is μ. In fact, for a normal distribution, mean = median = mode.
  - The median of a uniform distribution in the interval [*a*, *b*] is (*a* + *b*) / 2, which is also the mean.
- The median of a Cauchy distribution with location parameter x_0 and scale parameter *y* is x_0, the location parameter.
- The median of an exponential distribution with rate parameter λ is the natural logarithm of 2 divided by the rate parameter: $\lambda^{-1} \ln 2$.
- The median of a Weibull distribution with shape parameter *k* and scale parameter λ is $\lambda(\ln 2)^{1/k}$.

4.11.4 **Medians in descriptive statistics**



Comparison of mean, median and mode of two log-normal distributions with different skewness.

The median is used primarily for skewed distributions, which it summarizes differently than the arithmetic mean. Consider the multiset { 1, 2, 2, 2, 3, 14 }. The median is 2 in this case, (as is the mode), and it might be seen as a better indication of central tendency (less susceptible to the exceptionally large value in data) than the arithmetic mean of 4.

Calculation of medians is a popular technique in summary statistics and summarizing statistical data, since it is simple to understand and easy to calculate, while also giving a measure that is more robust in the presence of outlier values than is the mean.

4.11.5 **Medians for populations**

4.11.5.1**An optimality property**

The *mean absolute error* of a real variable *c* with respect to the random variable *X* is

$$E(|X - c|)$$

Provided that the probability distribution of *X* is such that the above expectation exists, then *m* is a median of *X* if and only if *m* is a minimizer of the mean absolute error with respect to *X*. In particular, *m* is a sample median if and only if *m* minimizes the arithmetic mean of the absolute deviations.

4.11.5.2 **Unimodal distributions**

It can be shown for a unimodal distribution that the median $\tilde{X}$ and the mean $\bar{X}$ lie within $(3/5)^{1/2}$ standard deviations of each other. In symbols,

$$\frac{\left|\tilde{X} - \bar{X}\right|}{\sigma} \leq (3/5)^{1/2}$$

where |.| is the absolute value.

A similar relation holds between the median and the mode: they lie within $3^{1/2}$ standard deviations of each other:

$$\frac{\left|\tilde{X} - \text{mode}\right|}{\sigma} \leq 3^{1/2}.$$

4.11.5.3 **An inequality relating means and medians**

If the distribution has finite variance, then the distance between the median and the mean is bounded by one standard deviation.

This bound was proved by Mallows, who used Jensen's inequality twice, as follows. We have

$$\begin{aligned} |\mu - m| = |\mathrm{E}(X - m)| &\leq \mathrm{E}\left(|X - m|\right) \\ &\leq \mathrm{E}\left(|X - \mu|\right) \\ &\leq \sqrt{\mathrm{E}((X - \mu)^2)} = \sigma. \end{aligned}$$

The first and third inequalities come from Jensen's inequality applied to the absolute-value function and the square function, which are each convex. The second inequality comes from the fact that a median minimizes the absolute deviation function

$$a \mapsto \mathrm{E}(|X - a|).$$

This proof can easily be generalized to obtain a multivariate version of the inequality, as follows:

$$\begin{aligned} \|\mu - m\| = \|\mathrm{E}(X - m)\| &\leq \mathrm{E}\|X - m\| \\ &\leq \mathrm{E}(\|X - \mu\|) \\ &\leq \sqrt{\mathrm{E}(\|X - \mu\|^2)} = \sqrt{\mathrm{trace}(\mathrm{var}(X))} \end{aligned}$$

where *m* is a spatial median, that is, a minimizer of the function $a \mapsto \mathrm{E}(\|X - a\|)$. The spatial median is unique when the data-set's dimension is two or more. An alternative proof uses the one-sided Chebyshev inequality; it appears in an inequality on location and scale parameters.

4.11.6 **Jensen's inequality for medians**

Jensen's inequality states that for any random variable *x* with a finite expectation *E*(*x*) and for any convex function *f*

$$f(E(x)) \leq E(f(x))$$

It has been shown that if *x* is a real variable with a unique median *m* and *f* is a C function then

$$f(m) \leq \mathrm{Median}(f(x))$$

A C function is a real valued function, defined on the set of real numbers *R*, with the property that for any real *t*

$$f^{-1}((-\infty, t]) = \{x \in R | f(x) \leq t\}$$

is a closed interval, a singleton or an empty set.

4.11.7 **Medians for samples**

4.11.7.1**The sample median**

4.11.7.2 ***Efficient computation of the sample median***

Even though sorting *n* items requires *O*(*n* log *n*) operations, selection algorithms can compute the k^{th}-smallest of *n* items (e.g., the median) with only O(*n*) operations.

4.11.7.3 ***Easy explanation of the sample median***

In individual series (if number of observation is very low) first one must arrange all the observations in ascending order. Then count(*n*) is the total number of observation in given data.

If ***n* is odd** then Median (*M*) = value of ((*n* + 1)/2)th item term.

If ***n* is even** then Median (*M*) = value of [((*n*)/2)th item term + ((*n*)/2 + 1)th item term]/2

For an odd number of values

As an example, we will calculate the sample median for the following set of observations: 1, 5, 2, 8, 7.

Start by sorting the values: 1, 2, 5, 7, 8.

In this case, the median is 5 since it is the middle observation in the ordered list.

The median is the ((*n* + 1)/2)th item, where *n* is the number of values. For example, for the list {1, 2, 5, 7, 8}, we have *n* = 5, so the median is the ((5 + 1)/2)th item.

median = (6/2)th item
median = 3rd item
median = 5

For an even number of values

As an example, we will calculate the sample median for the following set of observations: 1, 6, 2, 8, 7, 2.

Start by sorting the values: 1, 2, 2, 6, 7, 8.

In this case, the arithmetic mean of the two middlemost terms is (2 + 6)/2 = 4. Therefore, the median is 4 since it is the arithmetic mean of the middle observations in the ordered list.

We also use this formula MEDIAN = {(*n* + 1)/2}th item . *n* = number of values

As above example 1, 2, 2, 6, 7, 8 *n* = 6 Median = {(6 + 1)/2}th item = 3.5th item. In this case, the median is average of the 3rd number and the next one (the fourth number). The median is (2 + 6)/2 which is 4.

4.11.7.4 ***Variance***

The distribution of both the sample mean and the sample median were determined by Laplace. The distribution of the sample median from a population with a density function $f(x)$is asymptotically normal with mean $m$and variance

$$\frac{1}{4nf(m)^2}$$

where $m$is the median value of distribution and $n$is the sample size. In practice this may be difficult to estimate as the density function is usually unknown.

These results have also been extended. It is now known that for the p-th quartile that the distribution of the sample p-th quartile is distributed normally around the p-th quartile with variance equal to

$$\frac{p(1-p)}{nf(x_p)^2}$$

where $f(x_p)$is the value of the distribution at the p-th quartile.

Estimation of variance from sample data

The value of $(2f(x))^{-2}$—the asymptotic value of $n^{-\frac{1}{2}}(\nu - m)$where $\nu$is the population median—has been studied by several authors. The standard 'delete one' jackknife method produces inconsistent results. An alternative—the 'delete k' method—where $k$grows with the sample size has been shown to be asymptotically consistent. This method may be computationally expensive for large data sets. A bootstrap estimate is known to be consistent, but converges very slowly (order of $n^{-\frac{1}{4}}$). Other methods have been proposed but their behavior may differ between large and small samples.

Efficiency

The efficiency of the sample median, measured as the ratio of the variance of the mean to the variance of the median, depends on the sample size and on the underlying population distribution. For a sample of size $N = 2n + 1$from the normal distribution, the ratio is

$$\frac{4n}{\pi(2n+1)}$$

For large samples (as $n$tends to infinity) this ratio tends to $\frac{2}{\pi}$.

4.11.7.5 **Other estimators**

For univariate distributions that are *symmetric* about one median, the Hodges–Lehmann estimator is a robust and highly efficient estimator of the population median.

If data are represented by a statistical model specifying a particular family of probability distributions, then estimates of the median can be obtained by fitting that family of probability distributions to the data and calculating the theoretical median of the fitted distribution. Pareto interpolation is an application of this when the population is assumed to have a Pareto distribution.

4.11.8 **Coefficient of dispersion**

The coefficient of dispersion (CD) is defined as the ratio of the average absolute deviation from the median to the median of the data. It is a statistical measure used by the states of Iowa, New York and South Dakota in estimating dues taxes. In symbols

$$CD = \frac{1}{n}\frac{\sum |m - x|}{m}$$

where n is the sample size, m is the sample median and x is a variate. The sum is taken over the whole sample.

Confidence intervals for a two sample test where the sample sizes are large have been derived by Bonett and Seier This test assumes that both samples have the same median but differ in the dispersion around it. The confidence interval (CI) is bounded inferiorly by

$$\exp\left[\log\left(\frac{t_a}{t_b}\right) - z_\alpha\left(var\left[\log\left(\frac{t_a}{t_b}\right)\right]\right)^{0.5}\right]$$

where t_j is the mean absolute deviation of the j^{th} sample, *var*() is the variance and z_α is the value from the normal distribution for the chosen value of α: for $\alpha = 0.05$, $z_\alpha = 1.96$. The following formulae are used in the derivation of these confidence intervals

$$var[\log(t_a)] = \frac{\left(\frac{s_a^2}{t_a^2} + \left(\frac{x_a - \bar{x}}{t_a}\right)^2 - 1\right)}{n}$$
$$var[\log(t_a/t_b)] = var[\log(t_a)] + var[\log(t_b)] - 2r(var[\log(t_a)]var[\log(t_b)])^{0.5}$$

where r is the Pearson correlation coefficient between the squared deviation scores

$$d_{ia} = |x_{ia} - \bar{x}_a|_{\text{and}}\ d_{ib} = |x_{ib} - \bar{x}_b|$$

a and b here are constants equal to 1 and 2, x is a variate and s is the standard deviation of the sample.

4.11.9 **Multivariate median**

Previously, this article discussed the concept of a univariate median for a one-dimensional object (population, sample). When the dimension is two or higher, there are multiple concepts that extend the definition of the univariate median; each such multivariate median agrees with the univariate median when the dimension is exactly one. In higher dimensions, however, there are several multivariate medians.

4.11.9.1 **Marginal median**

The marginal median is defined for vectors defined with respect to a fixed set of coordinates. A marginal median is defined to be the vector whose components are univariate medians. The marginal median is easy to compute, and its properties were studied by Puri and Sen.

4.11.9.2 **Spatial median (L1 median)**

In a normed vector space of dimension two or greater, the "spatial median" minimizes the expected distance

$$a \mapsto \mathrm{E}(\|X - a\|),$$

where *X* and *a* are vectors, if this expectation has a finite minimum; another definition is better suited for general probability-distributions. The spatial median is unique when the data-set's dimension is two or more. It is a robust and highly efficient estimator of the population spatial-median (also called the "L1 median").

4.11.9.3 **Other multivariate medians**

An alternative to the spatial median is defined in a similar way, but based on a different loss function, and is called the Geometric median. The centerpoint is another generalization to higher dimensions that does not relate to a particular metric.

4.11.10**Other median-related concepts**

4.11.10.1 Pseudo-median

For univariate distributions that are *symmetric* about one median, the Hodges–Lehmann estimator is a robust and highly efficient estimator of the population median; for non-symmetric distributions, the Hodges–Lehmann estimator is a robust and highly efficient estimator of the population *pseudo-median*, which is the median of a symmetrized distribution and which is close to the population median. The Hodges–Lehmann estimator has been generalized to multivariate distributions.

4.11.10.2 Variants of regression

The Theil–Sen estimator is a method for robust linear regression based on finding medians of slopes.

4.11.10.3 Median filter

In the context of image processing of monochrome raster images there is a type of noise, known as the salt and pepper noise, when each pixel independently becomes black (with some small probability) or white (with some small probability), and is unchanged otherwise (with the probability close to 1). An image constructed of median values of neighborhoods (like 3×3 square) can effectively reduce noise in this case.

4.11.10.4 Cluster analysis

In cluster analysis, the k-medians clustering algorithm provides a way of defining clusters, in which the criterion of maximising the distance between cluster-means that is used in k-means clustering, is replaced by maximising the distance between cluster-medians.

4.11.10.5 Median-Median Line

This is a method of robust regression. The idea dates back to Wald in 1940 who suggested dividing a set of bivariate data into two halves depending on the value of the independent parameter x: a left half with values less than the median and a right half with values greater than the median. He suggested taking the means of the dependent $y$and independent $x$variables of the left and the right halves and estimating the slope of the line joining these two points. The line could then be adjusted to fit the majority of the points in the data set.

Nair and Shrivastava in 1942 suggested a similar idea but instead advocated dividing the sample into three equal parts before calculating the means of the subsamples. Brown and Mood in 1951 proposed the idea of using the medians of two subsamples rather the means. Tukey combined these ideas and recommended dividing the sample into three equal size subsamples and estimating the line based on the medians of the subsamples.

4.11.11 Median-unbiased estimators

Any *mean*-unbiased estimator minimizes the risk (expected loss) with respect to the squared-error loss function, as observed by Gauss. A *median*-unbiased estimator minimizes the risk with respect to the absolute-deviation loss function, as observed by Laplace. Other loss functions are used in statistical theory, particularly in robust statistics.

The theory of median-unbiased estimators was revived by George W. Brown in 1947:

An estimate of a one-dimensional parameter θ will be said to be median-unbiased if, for fixed θ, the median of the distribution of the estimate is at the value θ; i.e., the estimate underestimates just as often as it overestimates. This requirement seems for most purposes to accomplish as much as the mean-unbiased requirement and has the additional property that it is invariant under one-to-one transformation. [page 584]

Further properties of median-unbiased estimators have been reported. In particular, median-unbiased estimators exist in cases where mean-unbiased and maximum-likelihood estimators do not exist. Median-unbiased estimators are invariant under one-to-one transformations.

4.11.12 History

The idea of the median originated in Edward Wright's book on navigation (*Certaine Errors in Navigation*) in 1599 in a section concerning the determination of location with a compass. Wright felt that this value was the most likely to be the correct value in a series of observations.

In 1757, Roger Joseph Boscovich developed a regression method based on the L1 norm and therefore implicitly on the median.

The distribution of both the sample mean and the sample median were determined by Laplace in the early 1800s.

Antoine Augustin Cournot in 1843 was the first to use the term *median* (*valeur médiane*) for the value that divides a probability distribution into two equal halves. Gustav Theodor Fechner used the median (*Centralwerth*) in sociological and psychological phenomena. It had earlier been used only in astronomy and related fields. Gustav Fechner popularized the median into the formal analysis of data, although it had been used previously by Laplace.

Francis Galton used the English term *median* in 1881, having earlier used the terms *middle-most value* in 1869 and the *medium* in 1880.

4.12 **Mode (statistics)**

The **mode** is the value that appears most often in a set of data. The mode of a discrete probability distribution is the value x at which its probability mass function takes its maximum value. In other words, it is the value that is most likely to be sampled. The mode of a continuous probability distribution is the value x at which its probability density function has its maximum value, so, informally speaking, the mode is at the peak.

Like the statistical mean and median, the mode is a way of expressing, in a single number, important information about a random variable or a population. The numerical value of the mode is the same as that of the mean and median in a normal distribution, and it may be very different in highly skewed distributions.

The mode is not necessarily unique, since the probability mass function or probability density function may take the same maximum value at several points x_1, x_2, etc. The most extreme case occurs in uniform distributions, where all values occur equally frequently.

The above definition tells us that only *global maxima* are modes. Slightly confusingly, when a probability density function has multiple local maxima it is common to refer to all of the local maxima as modes of the distribution. Such a continuous distribution is called multimodal (as opposed to unimodal).

In symmetric unimodal distributions, such as the normal (or Gaussian) distribution (the distribution whose density function, when graphed, gives the famous "bell curve"), the mean (if defined), median and mode all coincide. For samples, if it is known that they are drawn from a symmetric distribution, the sample mean can be used as an estimate of the population mode.

4.12.1Mode of a sample

The mode of a sample is the element that occurs most often in the collection. For example, the mode of the sample [1, 3, 6, 6, 6, 6, 7, 7, 12, 12, 17] is 6. Given the list of data [1, 1, 2, 4, 4] the mode is not unique - the dataset may be said to be bimodal, while a set with more than two modes may be described as multimodal.

For a sample from a continuous distribution, such as [0.935..., 1.211..., 2.430..., 3.668..., 3.874...], the concept is unusable in its raw form, since no two values will be exactly the same, so each value will occur precisely once. In order to estimate the mode, the usual practice is to discretize the data by assigning frequency values to intervals of equal distance, as for making a histogram, effectively replacing the values by the midpoints of the intervals they are assigned to. The mode is then the value where the histogram reaches its peak. For small or middle-sized samples the outcome of this procedure is sensitive to the choice of interval width if chosen too narrow or too wide; typically one should have a sizable fraction of the data concentrated in a relatively small number of intervals (5 to 10), while the fraction of the data falling outside these intervals is also sizable. An alternate approach is kernel density estimation, which essentially blurs point samples to produce a continuous estimate of the probability density function which can provide an estimate of the mode.

4.12.2 Comparison of mean, median and mode

Comparison of common averages of values { 1, 2, 2, 3, 4, 7, 9 }

Type	Description	Example	Result
Arithmetic mean	Sum of values of a data set divided by number of values: $\bar{x}=\frac{1}{n}\sum_{i=1}^{n} x_i$	(1+2+2+3+4+7+9) / 7	**4**
Median	Middle value separating the greater and lesser halves of a data set	1, 2, 2, **3**, 4, 7, 9	**3**
Mode	Most frequent value in a data set	1, **2**, **2**, 3, 4, 7, 9	**2**

4.12. 3 Use

Unlike mean and median, the concept of mode also makes sense for "nominal data" (i.e., not consisting of numerical values in the case of mean, or even of ordered values in the case of median). For example, taking a sample of Korean family names, one might find that "Kim" occurs more often than any other name. Then "Kim" would be the mode of the sample. In any voting system where a plurality determines victory, a single modal value determines the victor, while a multi-modal outcome would require some tie-breaking procedure to take place.

Unlike median, the concept of mean makes sense for any random variable assuming values from a vector space, including the real numbers (a one-dimensional vector space) and the integers (which can be considered embedded in the reals). For example, a distribution of points in the plane will typically have a mean and a mode, but the concept of median does not apply. The median makes sense when there is a linear order on the possible values. Generalizations of the concept of median to higher-dimensional spaces are the geometric median and the centerpoint.

4.12.4 Uniqueness and definedness

For the remainder, the assumption is that we have (a sample of) a real-valued random variable.

For some probability distributions, the expected value may be infinite or undefined, but if defined, it is unique. The mean of a (finite) sample is always defined. The median is the value such that the fractions not exceeding it and not falling below it are both at least 1/2. It is not necessarily unique, but never infinite or totally undefined. For a data sample it is the "halfway" value when the list of values is ordered in increasing value, where usually for a list of even length the numerical average is taken of the two values closest to "halfway". Finally, as said before, the mode is not necessarily unique. Certain pathological distributions (for example, the Cantor distribution) have no defined mode at all. For a finite data sample, the mode is one (or more) of the values in the sample.

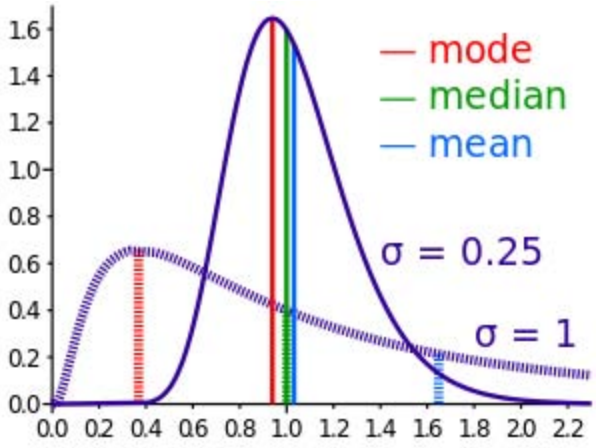
4.12.5 Properties

Assuming definedness, and for simplicity uniqueness, the following are some of the most interesting properties.

- All three measures have the following property: If the random variable (or each value from the sample) is subjected to the linear or affine transformation which replaces X by $aX+b$, so are the mean, median and mode.
- However, if there is an arbitrary monotonic transformation, only the median follows; for example, if X is replaced by $\exp(X)$, the median changes from m to $\exp(m)$ but the mean and mode won't.
- Except for extremely small samples, the mode is insensitive to "outliers" (such as occasional, rare, false experimental readings). The median is also very robust in the presence of outliers, while the mean is rather sensitive.
- In continuous unimodal distributions the median lies, as a rule of thumb, between the mean and the mode, about one third of the way going from mean to mode. In a formula, median ≈ (2 × mean + mode)/3. This rule, due to Karl Pearson, often applies to slightly non-symmetric distributions that resemble a normal distribution, but it is not always true and in general the three statistics can appear in any order.
- For unimodal distributions, the mode is within $\sqrt{3}$ standard deviations of the mean, and the root mean square deviation about the mode is between the standard deviation and twice the standard deviation.

4.12.6 Example for a skewed distribution

An example of a skewed distribution is personal wealth: Few people are very rich, but among those some are extremely rich. However, many are rather poor.



Comparison of mean, median and mode of two log-normal distributions with different skewness.

A well-known class of distributions that can be arbitrarily skewed is given by the log-normal distribution. It is obtained by transforming a random variable X having a normal distribution into random variable $Y = e^X$. Then the logarithm of random variable Y is normally distributed, hence the name.

Taking the mean μ of X to be 0, the median of Y will be 1, independent of the standard deviation σ of X. This is so because X has a symmetric distribution, so its median is also 0. The transformation from X to Y is monotonic, and so we find the median $e^0 = 1$ for Y.

When X has standard deviation $\sigma = 0.25$, the distribution of Y is weakly skewed. Using formulas for the log-normal distribution, we find:

$$\begin{array}{rlll} \text{mean} & = e^{\mu+\sigma^2/2} & = e^{0+0.25^2/2} & \approx 1.032 \\ \text{mode} & = e^{\mu-\sigma^2} & = e^{0-0.25^2} & \approx 0.939 \\ \text{median} & = e^{\mu} & = e^{0} & = 1 \end{array}$$

Indeed, the median is about one third on the way from mean to mode.

When X has a larger standard deviation, $\sigma = 1$, the distribution of Y is strongly skewed. Now

$$\begin{array}{rlll} \text{mean} & = e^{\mu+\sigma^2/2} & = e^{0+1^2/2} & \approx 1.649 \\ \text{mode} & = e^{\mu-\sigma^2} & = e^{0-1^2} & \approx 0.368 \\ \text{median} & = e^{\mu} & = e^{0} & = 1 \end{array}$$

Here, Pearson's rule of thumb fails.

4.12.7 van Zwet condition

Van Zwet derived an inequality which provides sufficient conditions for this inequality to hold. The inequality

$$\text{Mode} \leq \text{Median} \leq \text{Mean}$$

holds if

$$F(\text{Median} - x) + F(\text{Median} + x) \geq 1$$

for all x where F() is the cumulative distribution function of the distribution.

4.12.8 Unimodal distributions

The difference between the mean and the mode in a unimodal distribution is bounded by $3^{1/2}$. In symbols

$$\frac{|\text{mean} - \text{mode}|}{\text{standard deviation}} \leq \sqrt{3}$$

where || is the absolute value. Incidentally this formula is also the Pearson mode or first skewness coefficient, and shows that this statistic for a unimodal distribution is bounded by $3^{1/2}$.

The difference between the mode and the median has the same bound. In symbols

$$\frac{|\text{median} - \text{mode}|}{\text{standard deviation}} \leq \sqrt{3}$$

4.12.9 Confidence interval for the mode with a single data point

It is a common but false belief that from a single observation *x* we can not gain information about the variability in the population and that consequently that finite length confidence intervals for mean and/or variance are impossible even in principle.

It is possible for an unknown unimodal distribution to estimate a confidence interval for the mode with a sample size of 1. This was first shown by Abbot and Rosenblatt and extended by Blachman and Machol. This confidence interval can be sharpened if the distribution can be assumed to be symmetrical. It is further possible to sharpen this interval if the distribution is normally distributed.

Let the confidence interval be 1 - α. Then the confidence intervals for the general, symmetric and normally distributed variates respectively are

$$X \pm (\frac{2}{\alpha} - 1)|X - \theta|$$
$$X \pm (\frac{1}{\alpha} - 1)|X - \theta|$$
$$X \pm (\frac{0.484}{\alpha} - 1)|X - \theta|$$

where X is the variate, θ is the mode and || is the absolute value.

These estimates are conservative. The confidence intervals for the mode at the 90% level given by these estimators are $X \pm 19 | X - \theta |$, $X \pm 9 | X - \theta |$ and $X \pm 5.84 | X - \theta |$ for the general, symmetric and normally distributed variates respectively. The 95% confidence interval for a normally distributed variate is given by $X \pm 10.7 | X - \theta |$. It may be worth noting that the mean and the mode coincide if the variates are normally distributed.

The 95% bound for a normally distributed variate has been improved and is now known to be $X \pm 9.68 | X - \theta |$ The bound for a 99% confidence interval is $X \pm 48.39 | X - \theta'|$

Note

Machol has shown that that given a known density symmetrical about 0 that given a single sample value (x) that the 90% confidence intervals of population mean are

$$x \pm 5|x - \nu|$$

where *v* is the population median.

If the precise form of the distribution is not known but it is known to be symmetrical about zero then we have

$$P(X - k|X - a| \leq \mu \leq X + k|X - a|) \geq 1 - \frac{1}{1 + k}$$

where *X* is the variate, μ is the population mean and *a* and *k* are arbitrary real numbers.

It is also possible to estimate a confidence interval for the standard deviation from a single observation if the distribution is symmetrical about 0. For a normal distribution the with an unknown variance and a single data point (*X*) the 90%, 95% and 99% confidence intervals for the standard deviation are [0, 8|*X*|], [0, 17|*X*|] and [0, 70|*X*|]. These intervals may be shorted if the mean is known to be bounded by a multiple of the standard deviation.

If the distribution is known to be normal then it is possible to estimate a confidence interval for both the mean and variance from a simple value. The 90% confidence intervals are

$$X - 23.3|X| \leq \mu \leq X + 23.3|X|$$
$$\sigma \leq 10|X|$$

The confidence intervals can be estimated for any chosen range.

This method is not limited to the normal distribution but can be used with any known distribution.

4.12.10 Statistical tests

These estimators have been used to create hypothesis tests for simple samples from normal or symmetrical unimodal distributions. Let the distribution have an assumed mean (μ_0). The null hypothesis is that the assumed mean of the distribution lies within the confidence interval of the sample mean (*m*). The null hypothesis is accepted if

$$\mu_0 < \frac{x + m}{2} \pm k|x - m|$$

where *x* is the value of the sample and *k* is a constant. The null hypothesis is rejected if

$$\mu_0 > \frac{x + m}{2} \pm k|x - m|$$

The value of *k* depends on the choice of confidence interval and the nature of the assumed distribution.

If the distribution is assumed or is known to be normal then the values of *k* for the 50%, 66.6%, 75%, 80%, 90%, 95% and 99% confidence intervals are 0.50, 1.26, 1.80, 2.31, 4.79, 9.66 and 48.39 respectively.

If the distribution is assumed or known to be unimodal and symmetrical but not normal then the values of *k* for the 50%, 66.6%, 75%, 80%, 90%, 95% and 99% confidence intervals are 0.50, 1.87, 2.91, 3.94, 8.97, 18.99, 99.00 respectively.

To see how this test works we assume or know *a priori* that the population from which the sample is drawn has a mean of μ_0 and that the population has a symmetrical unimodal distribution - a class that includes the normal distribution. We wish to know if the mean estimated from the sample is representative of the population at a pre chosen level of confidence.

Assume that the distribution is normal and let the confidence interval be 95%. Then $k = 9.66$.

Assuming that the sample is representative of the population, the sample mean (*m*) will then lie within the range determined from the formula:

$$\mu_0 < \frac{x+m}{2} \pm 9.66|x-m|$$

If subsequent sampling shows that the sample mean lies outside these parameters the sample mean is to be considered to differ significantly from the population mean.

Review Questions

1. Define the Central Tendency?
2. Explain the Mean?
3. Explain the Median?
4. Explain the Mode?

Discussion Questions

Discuss the Measurements of central tendencies?

"The lesson content has been compiled from various sources in public domain including but not limited to the internet for the convenience of the users. The university has no proprietary right on the same."

EIILM UNIVERSITY
SIKKIM

Jorethang, District Namchi, Sikkim- 737121, India
www.eiilmuniversity.ac.in