CONTINUOUS DISTRIBUTIONS

6.1 **NORMAL DISTRIBUTION: In probability theory, the normal (or Gaussian) distribution is a very commonly occurring continuous probability distribution—a function that tells the probability that any real observation will fall between any two real limits or real numbers, as the curve approaches zero on either side.** Normal distributions are extremely important in statistics and are often used in the natural and social sciences for real-valued random variables whose distributions are not known.

The normal distribution is immensely useful because of the central limit theorem, which states that, under mild conditions, the mean of many random variables independently drawn from the same distribution is distributed approximately normally, irrespective of the form of the original distribution: physical quantities that are expected to be the sum of many independent processes (such as measurement errors) often have a distribution very close to the normal. Moreover, many results and methods (such as propagation of uncertainty and least squares parameter fitting) can be derived analytically in explicit form when the relevant variables are normally distributed.

The Gaussian distribution is sometimes informally called the **bell curve**. However, many other distributions are bell-shaped (such as Cauchy's, Student's, and logistic). The terms **Gaussian function** and **Gaussian bell curve** are also ambiguous because they sometimes refer to multiples of the normal distribution that cannot be directly interpreted in terms of probabilities.

The normal distribution is the only absolutely continuous distribution all of whose cumulants beyond the first two (i.e., other than the mean and variance) are zero. It is also the continuous distribution with the maximum entropy for a given mean and variance.

The normal distribution is a subclass of the elliptical distributions. The normal distribution is symmetric about its mean, and is non-zero over the entire real line. As such it may not be a suitable model for variables that are inherently positive or strongly skewed, such as the weight of a person or the price of a share. Such variables may be better described by other distributions, such as the log-normal distribution or the Pareto distribution.

The value of the normal distribution is practically zero when the value x lies more than a few standard deviations away from the mean. Therefore, it may not be an appropriate model when one expects a significant fraction of outliers—values that lie many standard deviations away from the mean — and least squares and other statistical inference methods that are optimal for normally distributed variables often become highly unreliable when applied to such data. In those cases, a more heavy-tailed distribution should be assumed and the appropriate robust statistical inference methods applied.

The Gaussian distribution belongs to the family of stable distributions which are the attractors of sums of independent, identically distributed distributions whether or not the mean or variance is finite. Except for the Gaussian which is a limiting case, all stable distributions have heavy tails and infinite variance.

Occurrence

The occurrence of normal distribution in practical problems can be loosely classified into four categories:

1. Exactly normal distributions;
2. Approximately normal laws, for example when such approximation is justified by the central limit theorem; and
3. Distributions modeled as normal – the normal distribution being the distribution with maximum entropy for a given mean and variance.
4. Regression problems – the normal distribution being found after systematic effects have been modeled sufficiently well.

Exact normality

Certain quantities in physics are distributed normally, as was first demonstrated by James Clerk Maxwell. Examples of such quantities are:

- Velocities of the molecules in the ideal gas. More generally, velocities of the particles in any system in thermodynamic equilibrium will have normal distribution, due to the maximum entropy principle.
- Probability density function of a ground state in a quantum harmonic oscillator.

Approximate normality

Approximately normal distributions occur in many situations, as explained by the central limit theorem. When the outcome is produced by many small effects

acting *additively and independently*, its distribution will be close to normal. The normal approximation will not be valid if the effects act multiplicatively (instead of additively), or if there is a single external influence that has a considerably larger magnitude than the rest of the effects.

- In counting problems, where the central limit theorem includes a discrete-to-continuum approximation and where infinitely divisible and decomposable distributions are involved, such as
  - Binomial random variables, associated with binary response variables;
  - Poisson random variables, associated with rare events;
- Thermal light has a Bose–Einstein distribution on very short time scales, and a normal distribution on longer timescales due to the central limit theorem.

Assumed normality

I can only recognize the occurrence of the normal curve – the Laplacian curve of errors – as a very abnormal phenomenon. It is roughly approximated to in certain distributions; for this reason, and on account for its beautiful simplicity, we may, perhaps, use it as a first approximation, particularly in theoretical investigations.

- In biology, the *logarithm* of various variables tend to have a normal distribution, that is, they tend to have a log-normal distribution (after separation on male/female subpopulations), with examples including:
  - Measures of size of living tissue (length, height, skin area, weight);
  - The *length* of *inert* appendages (hair, claws, nails, teeth) of biological specimens, *in the direction of growth*; presumably the thickness of tree bark also falls under this category;
  - Certain physiological measurements, such as blood pressure of adult humans.
- In finance, in particular the Black–Scholes model, changes in the *logarithm* of exchange rates, price indices, and stock market indices are assumed normal (these variables behave likecompound interest, not like simple interest, and so are multiplicative). Some mathematicians such as Benoît Mandelbrot have argued that log-Levy distributions, which possesses heavy tails would be a more appropriate model, in particular for the analysis for stock market crashes.
- Measurement errors in physical experiments are often modeled by a normal distribution. This use of a normal distribution does not imply that one is assuming the measurement errors are normally distributed, rather using the

normal distribution produces the most conservative predictions possible given only knowledge about the mean and variance of the errors.

- **In standardized testing, results can be made to have a normal distribution by either selecting the number and difficulty of questions (as in the IQ test) or transforming the raw test scores into "output" scores by fitting them to the normal distribution**. For example, the SAT's traditional range of 200–800 is based on a normal distribution with a mean of 500 and a standard deviation of 100.
- Many scores are derived from the normal distribution, including percentile ranks ("percentiles" or "quantiles"), normal curve equivalents, stanines,z-scores, and T-scores. Additionally, some behavioral statistical procedures assume that scores are normally distributed; for example, t-tests and ANOVAs. Bell curve grading assigns relative grades based on a normal distribution of scores.
- In hydrology the distribution of long duration river discharge or rainfall, e.g. monthly and yearly totals, is often thought to be practically normal according to the central limit theorem. The blue picture illustrates an example of fitting the normal distribution to ranked October rainfalls showing the 90% confidence belt based on the binomial distribution. The rainfall data are represented by plotting positions as part of the cumulative frequency analysis.

Produced normality

In regression analysis, lack of normality in residuals simply indicates that the model postulated is inadequate in accounting for the tendency in the data and needs to be augmented; in other words, normality in residuals can always be achieved given a properly constructed model.

6.2 SETS AND VENN DIAGRAMS: A Venn diagram or set diagram is a diagram that shows all possible logical relations between a finite collection of sets. Venn diagrams were conceived around 1880 by John Venn. They are used to teach elementary set theory, as well as illustrate simple set relationships in probability, logic,statistics, linguistics and computer science.

This example involves two sets, A and B, represented here as coloured circles. The orange circle, set A, represents all living creatures that are two-legged. The blue circle, set B, represents the living creatures that can fly. Each separate type of creature can be imagined as a point somewhere in the diagram. Living creatures that both can fly *and* have two legs—for example, parrots—are then in both sets,

so they correspond to points in the area where the blue and orange circles overlap. That area contains all such and only such living creatures.

Humans and penguins are bipedal, and so are then in the orange circle, but since they cannot fly they appear in the left part of the orange circle, where it does not overlap with the blue circle. Mosquitoes have six legs, and fly, so the point for mosquitoes is in the part of the blue circle that does not overlap with the orange one. Creatures that are not two-legged and cannot fly (for example, whales and spiders) would all be represented by points outside both circles.

The combined area of sets A and B is called the *union* of A and B, denoted by A ∪ B. The union in this case contains all living creatures that are either two-legged or that can fly (or both).

The area in both A and B, where the two sets overlap, is called the *intersection* of A and B, denoted by A ∩ B. For example, the intersection of the two sets is not empty, because there *are*points that represent creatures that are in *both* the orange and blue circles.

History

Venn diagrams were introduced in 1880 by John Venn (1834–1923) in a paper entitled *On the Diagrammatic and Mechanical Representation of Propositions and Reasonings* in the "Philosophical Magazine and Journal of Science", about the different ways to represent propositions by diagrams.[1][2] The use of these types of diagrams in formal logic, according to Ruskey and M. Weston, is "not an easy history to trace, but it is certain that the diagrams that are popularly associated with Venn, in fact, originated much earlier. They are rightly associated with Venn, however, because he comprehensively surveyed and formalized their usage, and was the first to generalize them".

Venn himself did not use the term "Venn diagram" and referred to his invention as "Eulerian Circles." For example, in the opening sentence of his 1880 article Venn writes, "Schemes of diagrammatic representation have been so familiarly introduced into logical treatises during the last century or so, that many readers, even those who have made no professional study of logic, may be supposed to be acquainted with the general nature and object of such devices. Of these schemes one only, viz. that commonly called 'Eulerian circles,' has met with any general acceptance..." The first to use the term "Venn diagram" was Clarence Irving Lewis in 1918, in his book "A Survey of Symbolic Logic".

Venn diagrams are very similar to Euler diagrams, which were invented by Leonhard Euler (1708–1783) in the 18th century. M. E. Baron has noted that Leibniz (1646–1716) in the 17th century produced similar diagrams before

Euler, but much of it was unpublished. She also observes even earlier Euler-like diagrams by Ramon Lull in the 13th Century.

In the 20th century, Venn diagrams were further developed. D.W. Henderson showed in 1963 that the existence of an *n*-Venn diagram with *n*-fold rotational symmetry implied that *n* was aprime number. He also showed that such symmetric Venn diagrams exist when *n* is 5 or 7. In 2002 Peter Hamburger found symmetric Venn diagrams for *n* = 11 and in 2003, Griggs, Killian, and Savage showed that symmetric Venn diagrams exist for all other primes. Thus rotationally symmetric Venn diagrams exist if and only if *n* is a prime number.

Venn diagrams and Euler diagrams were incorporated as part of instruction in set theory as part of the new math movement in the 1960s. Since then, they have also been adopted by other curriculum fields such as reading.

Overview

A Venn diagram is constructed with a collection of simple closed curves drawn in a plane. According to Lewis, the "principle of these diagrams is that classes [or *sets*] be represented by regions in such relation to one another that all the possible logical relations of these classes can be indicated in the same diagram. That is, the diagram initially leaves room for any possible relation of the classes, and the actual or given relation, can then be specified by indicating that some particular region is null or is not-null.

Venn diagrams normally comprise overlapping circles. The interior of the circle symbolically represents the elements of the set, while the exterior represents elements that are not members of the set. For instance, in a two-set Venn diagram, one circle may represent the group of all wooden objects, while another circle may represent the set of all tables. The overlapping area or *intersection* would then represent the set of all wooden tables. Shapes other than circles can be employed as shown below by Venn's own higher set diagrams. Venn diagrams do not generally contain information on the relative or absolute sizes (cardinality) of sets; i.e. they are schematic diagrams.

Venn diagrams are similar to Euler diagrams. However, a Venn diagram for *n* component sets must contain all 2^n hypothetically possible zones that correspond to some combination of inclusion or exclusion in each of the component sets. Euler diagrams contain only the actually possible zones in a given context. In Venn diagrams, a shaded zone may represent an empty zone, whereas in an Euler diagram the corresponding zone is missing from the diagram. For example, if one set represents *dairy products* and another *cheeses*, the Venn diagram contains a zone for cheeses that are not dairy products. Assuming that in

the context *cheese* means some type of dairy product, the Euler diagram has the cheese zone entirely contained within the dairy-product zone—there is no zone for (non-existent) non-dairy cheese. This means that as the number of contours increases, Euler diagrams are typically less visually complex than the equivalent Venn diagram, particularly if the number of non-empty intersections is small.