PROBABILITY

Staying Relevant

The central limit theorem remains a key concept in probability theory

by Julia Seaman, I. Elaine Allen and Samuel Zetumer

In the beginning, there was the central limit theorem (CLT). We use it every day as the basis for our statistical analyses. We base the reproducibility of our analyses on what the CLT tells us about our data, and we make inferences from the results of our hypothesis tests based on the CLT.

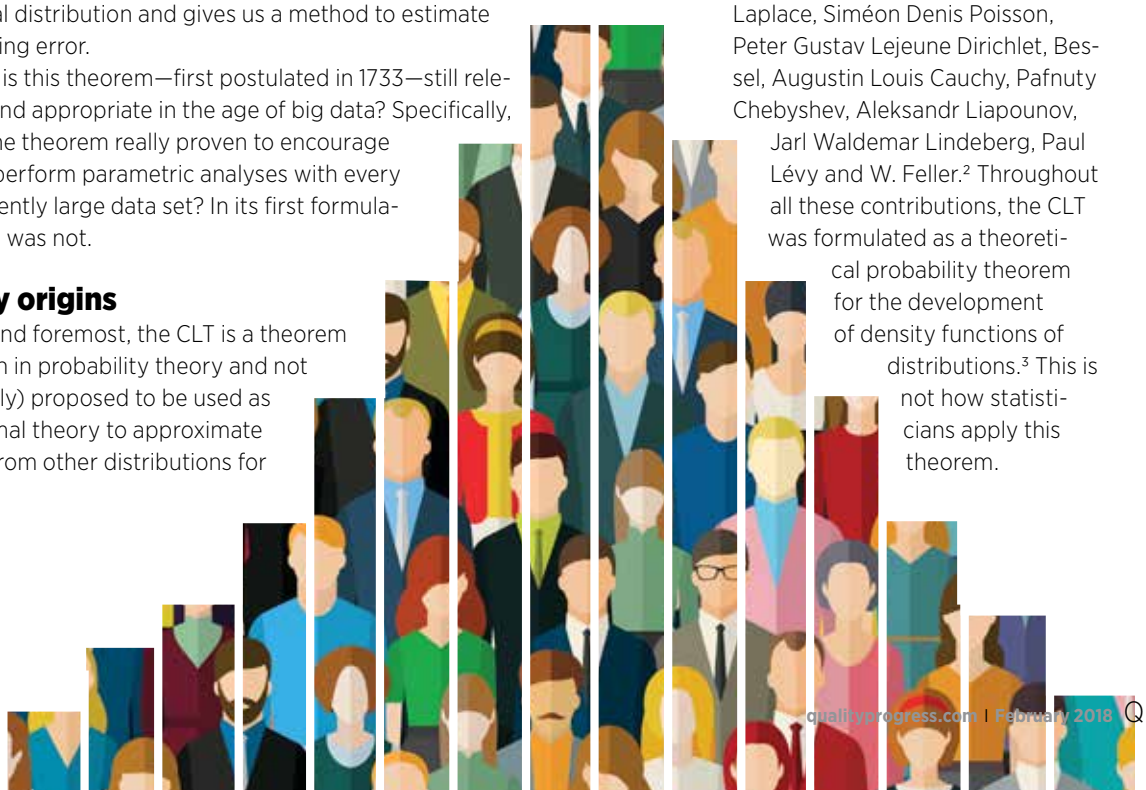
This theorem forms the basis for the majority of hypothesis testing and prediction models in statistics. In simplified terms, the CLT states that the sum of many different independent results tends toward a normal distribution and gives us a method to estimate sampling error.

But is this theorem—first postulated in 1733—still relevant and appropriate in the age of big data? Specifically, was the theorem really proven to encourage us to perform parametric analyses with every sufficiently large data set? In its first formulation, it was not.

Early origins

First and foremost, the CLT is a theorem proven in probability theory and not (initially) proposed to be used as a normal theory to approximate data from other distributions for statistical analysis. It was first proposed by Abraham de Moivre in 1733 when he discovered he could approximate binomial distribution probabilities from an integral of exp(-x2),[1] although he did not name this integral.

It wasn't until 1920 that George Polya gave the name to the theorem that we are now familiar with. There were several developments during the almost 200-year span from de Moivre to Polya to supplement the theorem to the results we use today. Over the centuries, the proof of the theorem included contributions by Pierre-Simon Laplace, Siméon Denis Poisson, Peter Gustav Lejeune Dirichlet, Bessel, Augustin Louis Cauchy, Pafnuty Chebyshev, Aleksandr Liapounov, Jarl Waldemar Lindeberg, Paul Lévy and W. Feller.[2] Throughout all these contributions, the CLT was formulated as a theoretical probability theorem for the development of density functions of distributions.[3] This is not how statisticians apply this theorem.



Statistical application

The theorem is derived under the assumption of an infinite population with observations that are independent, and identically distributed with constant mean and variance. Using basic calculus, it is not difficult to prove and is often included in high school curricula. Further, for this infinite population with mean, μ, and standard deviation, σ, there is the added assumption that for our sample to be normally distributed we must take sufficiently large random samples from the population with replacement. What de Moivre showed was that this will hold true regardless of the distribution of the source population.

FIGURE 1

Uniform distribution of the numbers 1–10 and mean 5.5

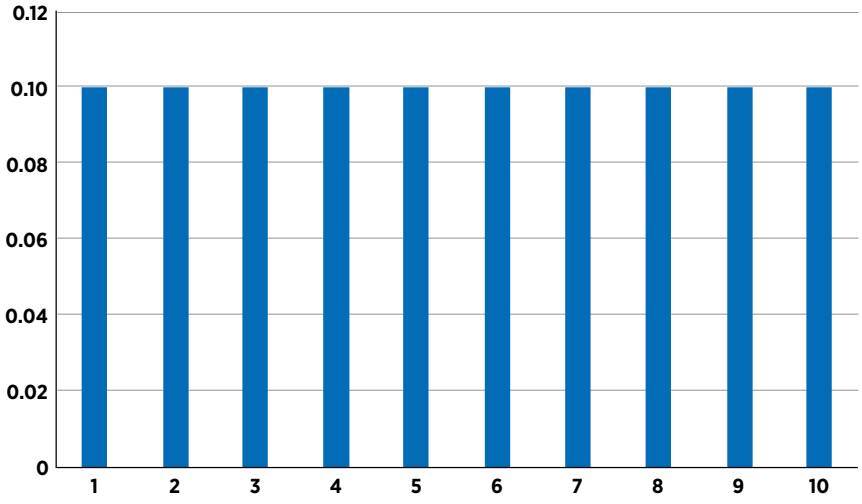


TABLE 1

Means of 3 from a uniform distribution of the numbers 1–10

Take samples of size 3			Average
1	2	3	2.0
1	2	4	2.3
1	2	5	2.7
1	2	6	3.0
1	2	7	3.3
1	2	8	3.7
1	2	9	4.0
1	2	10	4.3
...	...	...	...
6	8	9	7.7
6	8	10	8.0
6	9	10	8.3
7	8	9	8.0
7	8	10	8.3
7	9	10	8.7
8	9	10	9.0

In its most familiar form, this theorem does not apply to sampling from a finite population—for example, the number of factories an organization owns or the number of transit subway riders per day.[4] Two important modifications of the CLT were necessary before statisticians could apply the results to finite populations and sampling without replacement. Andrey Markov showed that the theorem can be relaxed for use with dependent sampling (without replacement) and Lévy showed that the same properties of the CLT with theoretical distribution can be applied to empirical distributions (that is, real data).[5]

In general, statisticians assume that whether the underlying distribution is normal or skewed, provided the sample size is sufficiently large (usually $n > 30$), the sample will be normal. If the population is already normal, the theorem holds true even for samples smaller than 30. In practice, this means we can use the normal probability model to quantify uncertainty when making inferences about a population mean based on the sample mean.

However, the essential component of the CLT is that it is referring to the distribution of our sample means approaching the normal distribution, and the mean of our sample means will be the same as the population mean, not a specific mean from one specific sample—as how the CLT is used today.

We are now analyzing large data sets from nonrandomized and from samples without replacement. The CLT, while very generalizable, was developed before the advent of computers and age of big data. Now, it's too easy to have too much data and therefore be magnitudes

greater than the lower limit of 30 samples. In these cases, the CLT may be valid, but other issues do arise when looking for the statistical signal in all the data that are not accounted for when the theorem and many parametric tests were developed.

Extreme example

Suppose we have a simple distribution of data in which each number is equally likely and the data are the first 10 numbers: 1, 2, 3, 4, 5, 6, 7, 8, 9 and 10. Note that the probability of taking a sample of size one from this group is 0.1, or 10%.

This is a small sample, but we can illustrate the power of the CLT by looking at the distribution of the means from all possible samples of size three. Table 1 shows the initial eight samples and the last seven samples for samples without replacement. Note that the means of each group of three are not necessarily close, varying from a mean of 2-9.

Sampling without replacement, there are 120 unique combinations of three numbers from the numbers 1-10. When all the combinations are counted together in a histogram, the average value of a sample of three approaches is a normal distribution (Figure 2) with an overall mean of 5.5. The overall mean of 5.5 is identical to the mean of the original distribution. While the original population contains 10 possible values, because our sampled population is more than 30, the CLT is applicable.

Sampling with replacement would increase the combinations to 220 and add both lower means (1 from the sample 1, 1, 1) and higher means (10 from the sample 10, 10, 10). The increased sample size shows a smoother normal

FIGURE 2

Histogram of the sample without replacement results showing a bell-shaped curve

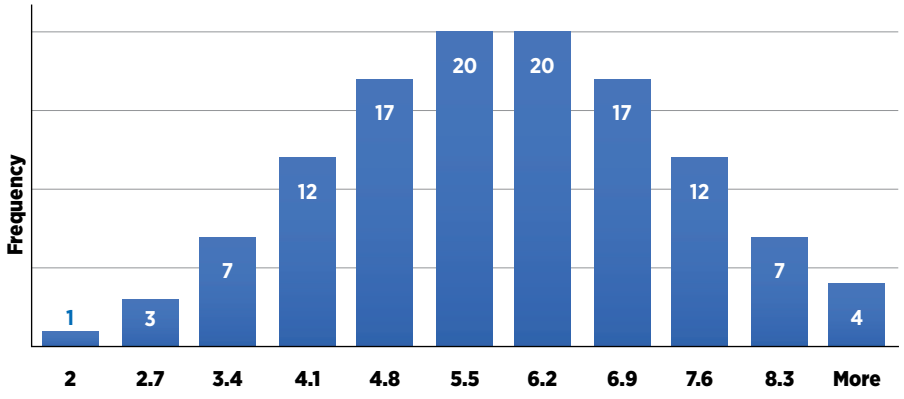
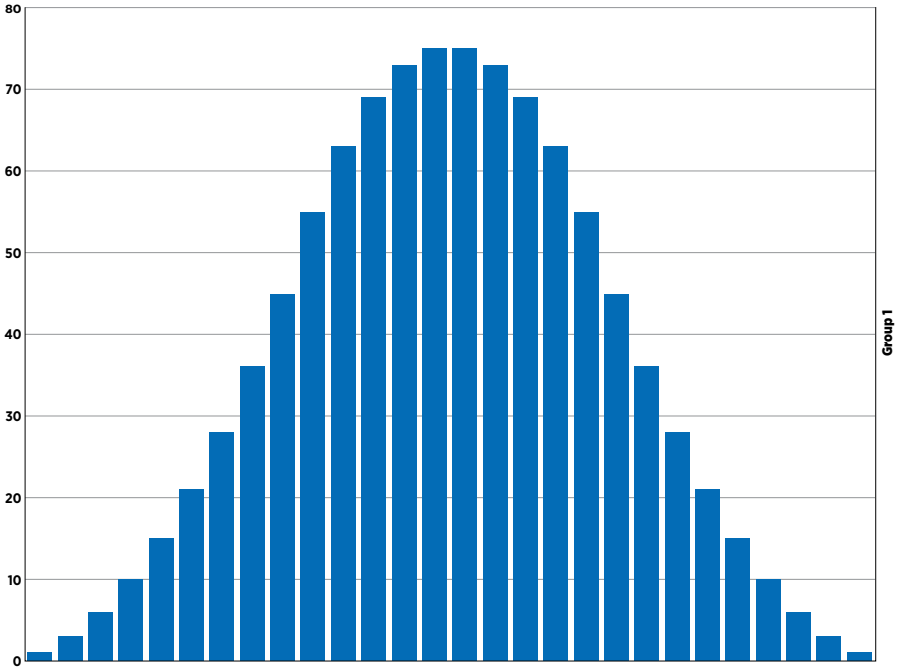


FIGURE 3

Histogram of the sample with replacement results showing a smoother bell-shaped curve



Editor's note: *The overall mean of the 1,000 means is 5.5, identical to the mean of the original distribution and the sample without replacement results.*

FIGURE 4

Resale home prices in 2015 in Singapore

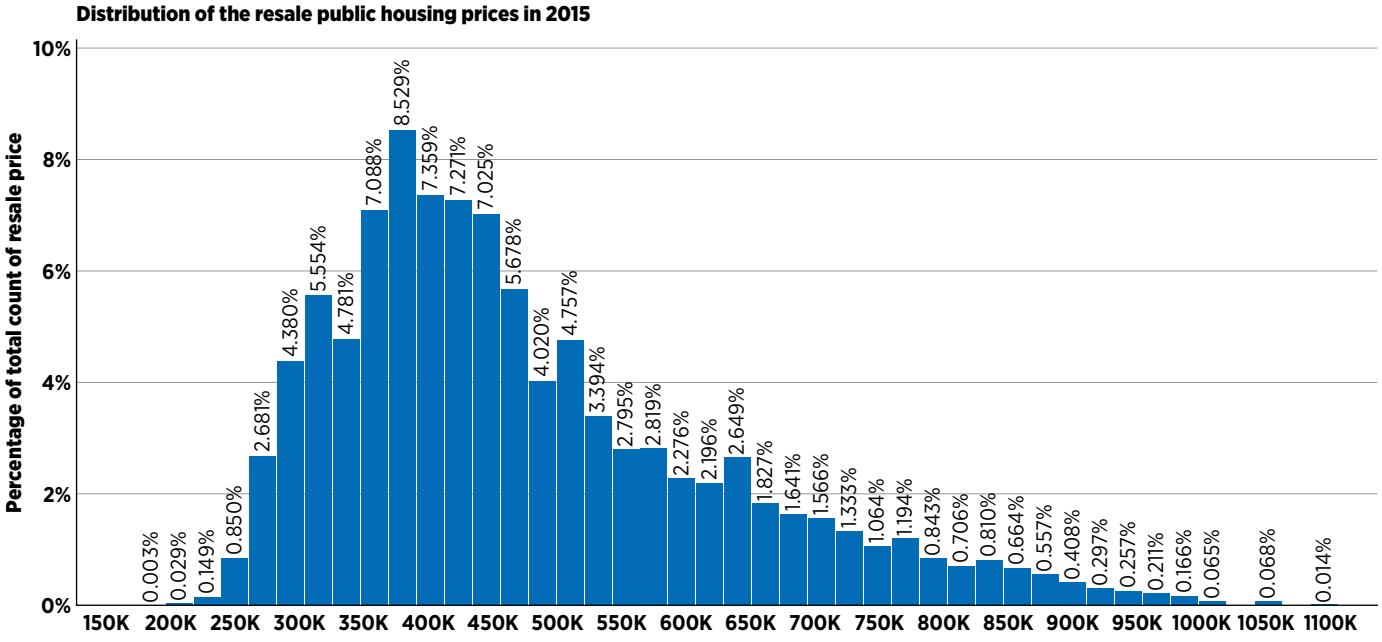
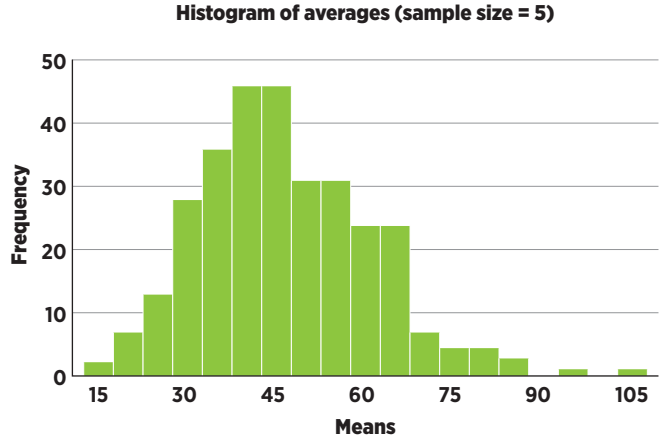
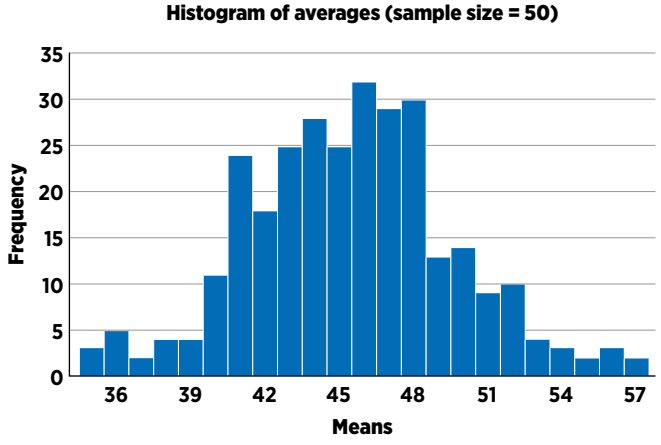


FIGURE 5

Sampling data approaching a normal distribution from the non-normal housing data





distribution curve than the previous histogram, as expected from the CLT (Figure 3, p. 55).

This extreme example may be unrealistic. Because one of the authors is a consultant, clients often share their data with the caveat that they know it is not normally distributed so parametric statistics based on the normal distribution should not be used. A good example of non-normally distributed data is housing prices. See Figure 4, which shows few homes are close to zero in cost and there are some expensive properties creating a highly skewed distribution (usually following a Weibull distribution).[6]

As discussed in a *New York Times* article, "...even when raw data does not fit a normal distribution, there is often a normal distribution lurking within it."[7,8]

It is not the raw data that form the basis for our use of statistics based on the normal distribution, but our reliance on the theorem to assure to us that the distribution of our sample mean will be normally distributed.

For samples of size five and size 50 from this distribution, the histograms in Figure 5 show much less skewness. Therefore, some parametric tests may be appropriate with this data set. When the sample is small and the data are skewed, for example, a test based on the median rather than the mean may be appropriate. The median always will represent the center of the distribution while the mean will be influenced (pulled in the direction) by the extreme skewness of the data.

Caveats and conclusions

If our sample size exceeds 30, can we always assume the CLT and use our parametric statistical results? Probably, but not always.

First, we should look at the data. If our key variable shows a strong bimodal distribution, using the normal distribution will mask the real differences in the two peaks in our data.

Next, if we are fitting models, look at the distributions of the independent and dependent variables and how they are related. If we are predicting ordinal or categorical variables, the normal distribution may not be appropriate regardless of the sample size.

Finally, with small samples, use parametric (based on the normal distribution) and nonparametric methods. Do they agree? Great! If not, look at your data more closely to understand these differences and which method assumptions are most accurate for the data. QP

REFERENCES

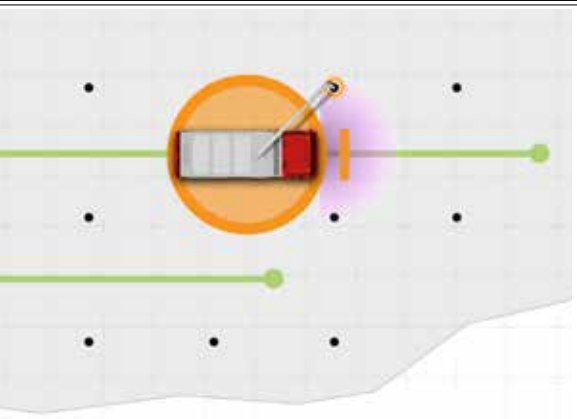
1. William J. Adams, *The Life and Times of the Central Limit Theorem*, second edition. American Math Society, 2009.
2. Max Mether, "A History of the Central Limit Theorem," *Specialized Applied Mathematics*, 2003, https://tinyurl.com/mether-clt-history.
3. Hans Fischer, *A History of the Central Limit Theorem: From Classical to Modern Probability Theory,* Springer, 2011.
4. D.R. Bellhouse, "The Central Limit Theorem Under Simple Random Sampling," *American Statistician*, Vol. 55, No. 4, 2001.
5. L. Le Cam, "The Central Limit Theorem Around 1935," *Statistical Science,* Vol. 1, No. 1, 1986, pp. 78-96.
6. Wikipedia, "ISSS608 2016-17 T1 Franky Eddy," https://tinyurl.com/wiki-franky-eddy.
7. Casey Dunn, "As 'Normal' as Rabbits' Weights and Dragons' Wings," *New York Times,* Sept. 23, 2013, https://tinyurl.com/nyt-casey-dunn.
8. Michelle Paret, "Explaining the Central Limit Theorem With Bunnies and Dragons," Minitab Blog, Oct. 15, 2013, https://tinyurl.com/minitab-blog-clt.

Julia E. Seaman is research director of the Quahog Research Group and a statistical consultant for the Babson Survey Research Group at Babson College in Wellesley, MA. She earned her doctorate in pharmaceutical chemistry and pharmacogenomics from the University of California, San Francisco.

I. Elaine Allen is professor of biostatistics at the University of California, San Francisco, and emeritus professor of statistics at Babson College. She is also director of the Babson Survey Research Group. She earned a doctorate in statistics from Cornell University in Ithaca, NY. Allen is a member of ASQ.

Samuel Zetumer is a medical degree candidate at the University of California, San Francisco, and a tutor and teaching assistant in biostatistics in the department of epidemiology and biostatistics. He earned a bachelor's degree in mathematics from Princeton University in New Jersey.

Marketplace



SOFTWARE

Mining platform includes blasting functionality

ASI Mining has collaborated with Enaex to develop semi-autonomous blasting functionality with ASI's autonomous command and control software, Mobius.

ASI's Mobius for Blasting application provides capability for teleoperation and autonomous navigation of blast vehicles, including mobile manufacturing unit and stemming vehicles. In addition, Mobius has the potential to coordinate drill and blasting, resulting in dynamically tailored blast processes based on actual "as-drilled" hole data, creating higher efficiency and increased fragmentation.

Steps were taken to ensure the autonomous blasting solutions meet all required safety, operational and availability standards, given the high risk of danger for workers and equipment.

This development is part of an ecosystem of teleoperation and autonomous units that will improve workers' safety by using technology to perform tasks in risky mine environments from a safe location.

www.asirobots.com | 866-881-2171

PROTECTIVE VENTS

Venting in potentially explosive environments

W. L. Gore & Associates' GORE PolyVent Ex+ is the latest addition to Gore's protective vents screw-in series and is certified according to explosion-proof safety standards, IECEx and ATEX. With these certifications, PolyVent Ex+ is allowed in areas with potentially explosive atmospheres caused by combustible gases or dust.

Materials selected for designing GORE PolyVent Ex+ were chosen to support the vent's long-lasting behavior in the field. The vent body, cap and membrane-sealing technology use nonflammable, stainless steel. The GORE membrane, made of 100% ePTFE, delivers performance for pressure-equalization customers, while achieving the highest flammability rating in its category. The silicone O-ring with a flammability resistance rating of UL 94 V-0 adds another layer of safety.

These materials, combined with the GORE PolyVent Ex+ construction, ensure flammability and corrosion resistance, and chemical robustness. The GORE membrane provides oleophobic and hydrophobic protection. With an airflow rate of 1,600 ml/min at 70 mbar and an ingress protection rating of IP68/IP69k, PolyVent Ex+ reliably protects enclosures up to 20l for a wide range of temperatures.

gore.com/protectivevents | 410-506-3526

POWER TAKE OFF

PTOs for demanding work conditions

Twin Disc's HP800, a hydraulically-actuated power take-off (PTO), is a middle horsepower range option for industrial applications. The HP800 has a maximum power rating of 800 hp at 1,800 rpm. It's ideal for driving pumps, grinders, crushers, dredgers, chippers, shredders and heavy-duty drills.

The key feature of the HP800 is the auxiliary drive pump towers 400 hp maximum capacity tower, or 450 hp maximum for both. They can rotate 0°, 45° and 90°, either clockwise or counter-clockwise, to allow for clearance in installations.

www.twindisc.com | sales@twindisc.com