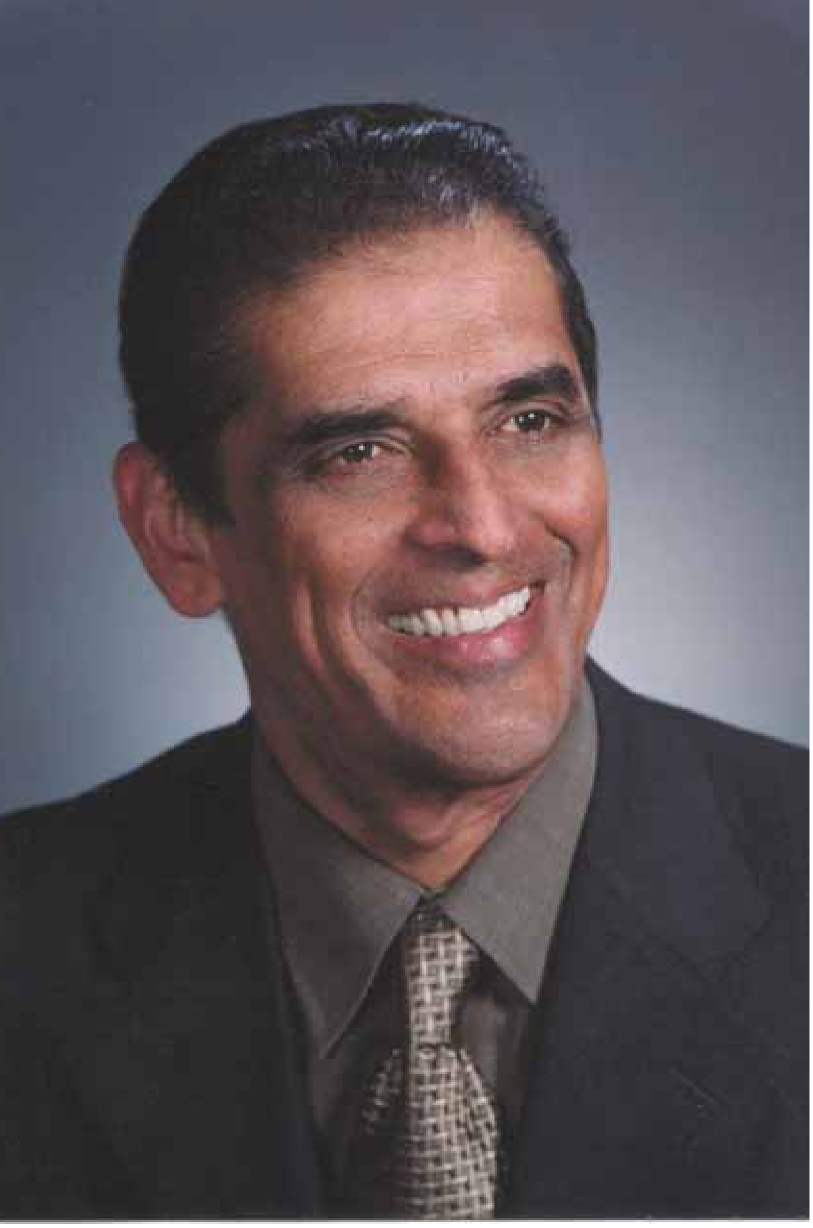


Statistical Reasoning vs. Magical Thinking

Shamanism as Statistical Knowledge: Is a Sample Size of 30 All You Need?

One of the most respected names in the marketing research industry weighs in against the notion that a sample size of 30 makes it possible to derive reliable statistical estimates. There is, he explains, no support for this, or any other, magical number.

Chuck Chakrapani, PhD, CMRP

Denis Diderot, the eminent encyclopedist, was visiting the Russian court just before the French Revolution. With his wit and charm, he began converting nobility to his atheistic ways. Alarmed by this, the czarina commissioned Euler, the famed mathematician, to debate Diderot. Told that Euler had found a mathematical proof for the existence of god, Diderot was hauled into court to debate the mathematician, who revealed his proof in all its gravity: "$(a + b^n)/n = x$. Therefore god exists. What is your response?" The brilliant Diderot had only a rudimentary knowledge of mathematics and didn't realize that the equation had nothing to do with the existence of god. He abruptly left the court and returned to France, much to the relief of the czarina.

As marketing researchers, we cannot afford to be as ignorant of formulas and numbers as Diderot was, and we try to become numerate. Some of us succeed, while others succumb to magical thinking, attributing supernatural powers to barely understood statistical statements, theorems and conclusions. My thinking on this subject was triggered by a discussion in which I said that small samples of 30 or even around 50 are inadequate to draw numeric conclusions about a relevant population, especially if the survey is of the type such as street intercepts or mall surveys. The person defending such research countered with the implied argument that, even in such conditions, a sample of 30 is good enough to draw numeric conclusions. The support for a sample size of 30 is assumed to come from the central limit theorem. Except that it doesn't. This got me reflecting on magical thinking in general and sample sizes in particular.

Where does it come from, the idea that there is this magical number 30, and if you have a sample of 30, you can derive reliable quantitative estimates from it? The central limit theorem states only that the mean of a sufficiently large number of independent random variables, each with finite mean and variance, will be approximately normally distributed. The theorem deliberately does not define what "large" means. If it could be proven that it is 30, or any other number for that matter, the theorem would have said so. But it does not.

Even more importantly, the theorem is based on theoretically perfect samples with replacement, not the samples we achieve in marketing research surveys. Where the curve generated by repeated samples of a given size converges to a normal curve would depend on the underlying distribution from which the sample is drawn and the variability with which it is associated. If the underlying distribution is perfectly normal and the sample is perfectly random with no non-response, coverage bias, or non-sampling error, it may approximate the normal curve quickly. If the underlying distribution is not normal or is skewed, we need larger samples. The theorem also says that, as the sample becomes *large*, the distribution of sample means becomes *approximately* normal (not *precisely* normal).

If the central limit theorem is silent about the meaning of "large," where does the shamanistic reverence for 30 as the sample size carrying mighty powers come from? It comes from artificial computer simulation experiments presented in introductory textbooks. These experiments take repeated idealized computer samples (assuming no error component) from a normal distribution, sometimes from skewed distributions.

But we know that many attributes in real life are not normally distributed. For example, consumer purchases follow a negative binomial distribution and not a normal distribution. Admission in maternity wards will likely follow Poisson distribution rather than a normal distribution. In a simulation exercise involving four different underlying distributions (normal, uniform, beta and gamma) carried out by Professor Murtaza Haider of the Ted Rogers School of Management, it took a sample of 4,500 (not 30) for the *t*-value to converge *precisely* to the *z*-values needed for a normal distribution. This is after assuming perfect random sampling, 100 per cent response rate, and no coverage error!

An obvious fact is that marketing research surveys differ markedly from the computer simulation exercises quoted in introductory textbooks in two crucial ways: First, in computer simulation, every sample is a perfect simple random sample. This is simply not possible in survey research. It is absurd to believe (magical thinking?) that our surveys, no matter how well they are conducted, achieve anything close to simple random sampling. The best we can hope for is that, in well-executed surveys, the results could approximate those generated by random samples.

Second, the computer simulations of the central limit theorem do not include non-coverage, non-response, or non-sampling errors. In research surveys these are perennial problems. For example, the average response rate in surveys is 12 per cent, as opposed to an assumed 100 per cent response rate in computer simulations!

Therefore, it is naïve to assume that, just because computer-drawn samples of 30 achieve approximate convergence in simulation exercises, any sample of 30 would work the same way in survey research when conditions differ markedly.

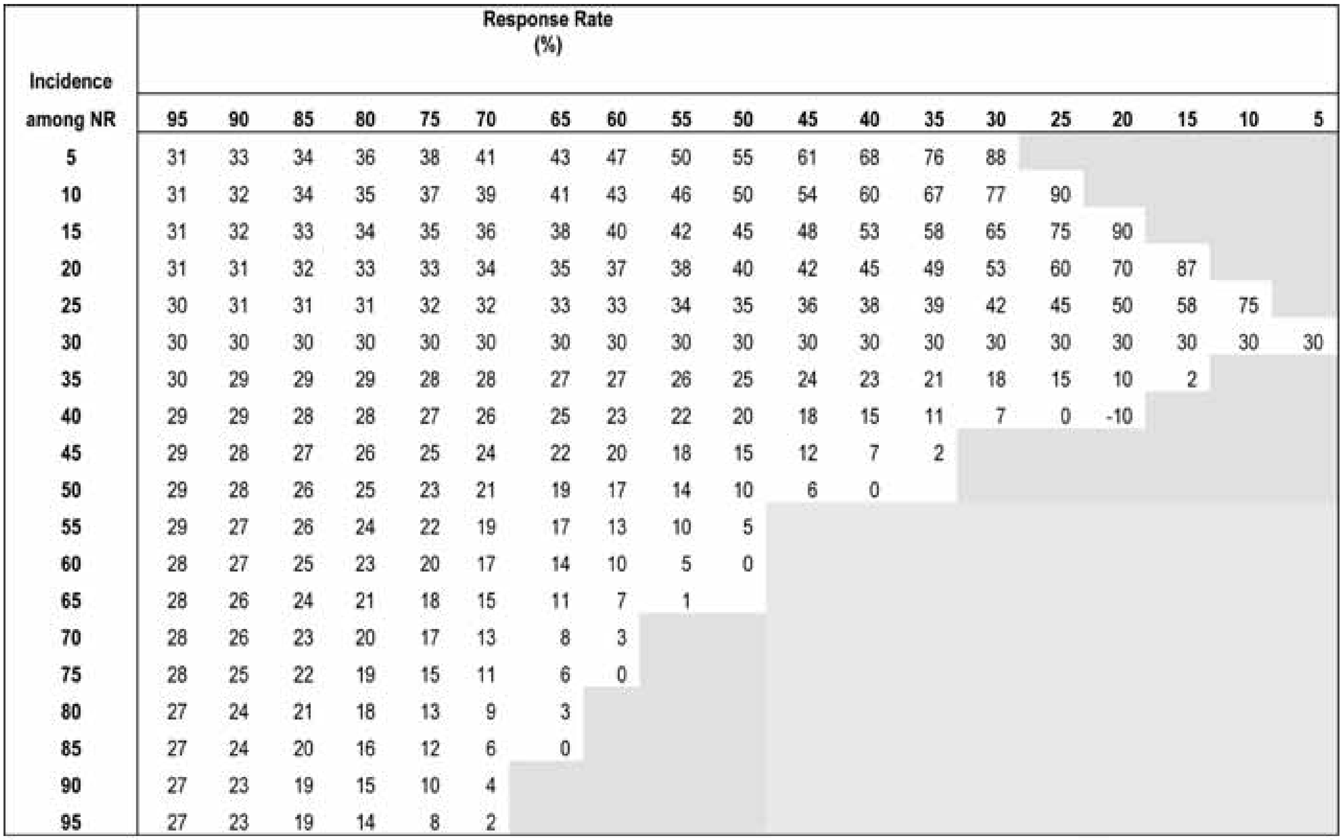
The central limit theorem is a very important theorem in statistics. It provides the basis for much of our sampling procedures. The fact that even small samples can converge to normality is interesting and has profound implications for marketing and social research. But it stretches credulity to take an inductive leap and believe that, therefore, the number 30 has magical properties and would work irrespective of the underlying distribution, irrespective of where and how sample is chosen, irrespective of clustering, irrespective of non-response, irrespective of non-randomness, and irrespective of other non-sampling errors that accompany marketing research studies. The central limit theorem simply does not say it, nor is there any empirical support for it.

Non-response is a reality in any marketing researcher's work life. The accompanying table shows what effect non-response can have on our results. Non-response is not taken into account in the "proof" offered by simulation exercises that appear in textbooks to illustrate that even a sample size of 30, under some conditions, could result in convergence.

As a matter of fact, there is no magic in a sample of 100 either (and I consider this number the approximate minimum for certain types of studies for control conditions). I use 100 because it is generally considered a reasonable minimum sample size by marketing researchers, based on their experience with several thousand studies over several decades. Also, at around this point, the *t*-test values begin to get much closer to the *z*-scores based on the normal curve.

All that the central limit theorem says is that "as the sample size becomes large … ." We can either apply statistical thinking and base our interpretation of what a "large number" might be in a given context, preferably, on empirical observations (subject to revision, should empirical results show otherwise), or succumb to a shallow interpretation of the theorem by attributing magical properties to some arbitrary number such as 30.

What is completely overlooked in this irrelevant invocation of the supposed sacredness and the might of the sample of 30 is the matter of the validity of the survey itself. I do not hold that 30 is an inadequate sample size in all contexts. For example, to assess the impact of a fertilizer on crops, it is possible to take just 10 homogeneous plots of land, divide each into two parts, apply the fertilizer to one part and not to the other part. The crop yields of these 10 split plots could potentially provide valid experimental results as to the efficacy of the fertilizer. Even in cases where



Incidence	Response Rate (%)																		
among NR	95	90	85	80	75	70	65	60	55	50	45	40	35	30	25	20	15	10	5
5	31	33	34	36	38	41	43	47	50	55	61	68	76	88					
10	31	32	34	35	37	39	41	43	46	50	54	60	67	77	90				
15	31	32	33	34	35	36	38	40	42	45	48	53	58	65	75	90			
20	31	31	32	33	33	34	35	37	38	40	42	45	49	53	60	70	87		
25	30	31	31	31	32	32	33	33	34	35	36	38	39	42	45	50	58	75	
30	30	30	30	30	30	30	30	30	30	30	30	30	30	30	30	30	30	30	30
35	30	29	29	29	28	28	27	27	26	25	24	23	21	18	15	10	2		
40	29	29	28	28	27	26	25	23	22	20	18	15	11	7	0	-10			
45	29	28	27	26	25	24	22	20	18	15	12	7	2						
50	29	28	26	25	23	21	19	17	14	10	6	0							
55	29	27	26	24	22	19	17	13	10	5									
60	28	27	25	23	20	17	14	10	5	0									
65	28	26	24	21	18	15	11	7	1										
70	28	26	23	20	17	13	8	3											
75	28	25	22	19	15	11	6	0											
80	27	24	21	18	13	9	3												
85	27	24	20	16	12	6	0												
90	27	23	19	15	10	4													
95	27	23	19	14	8	2													

The above table illustrates what could happen with non-response in a case in which the true incidence is 30%. Let us assume that the response rate is 50%. If the incidence rate among non-responders is 35%, then our sample could potentially show an incidence rate that is as low as 25%. (This is the value shown in a cell that intersects a 50% response rate and a 35% incidence rate among non-responders.)

our sample is otherwise small, even when it is less than 30, we can apply non-parametric tests or *t*-tests, as William Gossett famously did at Guinness Breweries. So the proper question should be "Is the sample size adequate for the intended purpose?" (In our hypothetical example, the purpose is to establish a quantitative estimate in street survey research, with all its attendant imperfections of sample selection.) For some problems, a sample size of 10 may be adequate; for others (such as data mining or text mining of social media that culls data from millions of online conversations), a sample of 10,000 may be considered small. One cannot decide on the sample size based on statistical formulas alone without considering the context.

Hardly any statistics books written by statisticians (as opposed to those by social scientists and business professors) say that 30 is an adequate sample size or state that the central limit theorem endorses a sample size of 30 in survey research contexts, where non-coverage and non-response are major issues. The samples referred to in the central limit theorem are pure random samples and not samples that are subject to coverage, non-response, and non-sampling errors.

Mathematical theorems are precisely worded for a reason. Change or ignore a couple of words in a theorem and ignore an assumption, and you change the meaning of it. Take the central limit theorem: Change "large number" to "a sample size of 30," change "approximately" to "exactly," and ignore the fact that the samples referred to in the theorem are error-free, and voila! We have transformed a sophisticated statistical theorem into street magic.

As Aldous Huxley said, "Facts are ventriloquists' dummies. Sitting on a wise man's knee they may be made to utter words of wisdom; elsewhere, they say nothing, or talk nonsense." It applies equally, if not more so, to statistical facts.

They say a little knowledge is a dangerous thing. And so it is.

Dr. Chuck Chakrapani, president (Toronto) of Léger Marketing and distinguished visiting professor at the Ted Rogers School of Management, is a fellow of the Royal Statistical Society, fellow of MRIA, and editor of Marketing Research. *He is the chief knowledge officer of Blackstone Group, Chicago, and a board member of Marketing Research Institute International. Chuck is the author of hundreds of articles and over a dozen books. His latest is a university-level text,* Business Statistics for Contemporary Decision Making (2010), *co-authored with Black and Castillo.*