Revista Colombiana de Estadística
No. 31 y No. 32

NORMAL APPROXIMATION WHEN SAMPLING FROM BOUNDED SYMMETRIC DISTRIBUTIONS

JAIRO CLAVIJO M.
Department of Mathematics
Universidad del Tolima
Ibagué, Colombia

DAVID OSPINA
Department of Mathematics and Statistics
Universidad Nacional de Colombia
Santafé de Bogotá, Colombia

KEY WORDS AND PHRASES: Bounded symmetric distribution; equivalence relation; Finucan's principle; kurtosis; Monte Carlo simulation; normal approximation; sample size; triangular distribution; Van Zweet's ordering.

ABSTRACT. This paper introduces a basic methodology which, based on simulations, determines the minimum sample size necessary to use the normal approximation when estimating means from populations with bounded symmetric distributions.

1. INTRODUCTION

The sample size is a crucial aspect of statistical research since it must be large enough to guarantee the prefixed confidence level for the estimations and as small as possible to minimize the costs.

The determination of sample size is a problem of difficult solution using analytic methods. Generally, researchers make use of the central limit theorem, the law of large numbers or the Chebychev inequality. However, in many cases the values obtained with these methods are large and the budget assigned to the study is too small to satisfy them. Financial restrictions are, usually, the main reason to look for new procedures.

In this article, some mathematical results, derived from well-known general concepts, as well as certain simulation techniques, are used to determine those minima

values that support the normal approximation in the estimation process of means for bounded symmetric distributions. The final results show the convenience of the methodology used which could be implemented with other problems of similar caracteristics.

2. Preliminary Theory

For theoretical convenience this paper has been oriented to bounded symmetric distributions which possess an unique extreme value –maximum or minimum– within an open interval (a, b).

Any probability distribution F whose density function f is continous, symmetric and defined on an interval of finite lenght $[a, b]$, will be called **Bounded Symmetric Distribution**.

As a basic theoretic concept to develop the task, the classic kurtosis has been chosen since many authors like Finucan, Chisom, Darlington and Meeden [1] consider it to be a definitive element in the determination of the distribution shape. Van Zweet [1] considers the kurtosis inherent to the nature of the symmetric distributions and has introduced a partial order within the family of such distributions which is directly related to the kurtosis values corresponding to each of them. This order is established in [1] as follows:

Definition 1. (Van Zweet's ordering) Given f and g symmetric distributions, it is said that $f \leq g$ if and only if $g^{-1}(f(x))$ is convex for $x > m_f$ where m_f is the symmetry point of f.

The symmetry of the functions implies in particular that $g^{-1}(f(x))$ is convex for $x > m_f$ if and only if it is concave for $x < m_f$.

Van Zweet himself demonstrated that the following order of increasing dominance is fullfiled for these distributions:

$$\cup\text{-Shaped} \leq \text{Uniform} \leq \text{Normal} \leq \text{Logistic} \leq \text{Double Exponential}.$$

Another result, also due to Van Zweet, that constitutes an important element for this paper, establishes that if $f \leq g$ then $\gamma(f) \leq \gamma(g)$, where γ represents the classic kurtosis, defined as the quotient between the fourth central moment and the square of the variance.

Finucan [1] showed that if f and g are symmetric distributions with 0 mean and equal variance and if the graph of $g(x) - f(x)$ has the trend "peak-trough-peak" when x increases, then $\gamma(f) \leq \gamma(g)$. This result is known as ***Finucan's Principle***. The goal is to combine Van Zweet's results with Finucan's Principle to obtain a family of distributions that can be used in the solution of the problem described above.

3. A Solution Plan

Since Finucan's Principle is applicable to symmetric distributions with mean 0, the problem for arbitrary distributions must be translated to centered distributions (at the origin) and it must be shown that the basic properties of such distributions are not affected by translations.

Theorem 1. *Let X be a continuous random variable with density function $f(x)$ defined on an interval $[a, b]$ and let Y be a random variable defined by $Y = X + l$ with density function $g(y)$, defined on $[a + l, b + l]$. Then $\mu_r(X) = \mu_r(Y)$ for $l \in \mathbb{R}$ and $r = 1, 2, 3, ...$, where μ_r represents the r^{th} central moment.*

Proof. $y = t(x) = x + l$ and $x = v(y) = y - l$ are continuous monotonic functions.

Then (Cfr. [2], page 277) $g(y)$ can be written as:

$$g(y) = f(v(y))|v'(y)| = f(v(y)) = f(y-l) \quad \text{for} \quad a+l \leq y \leq b+l$$

where $|v'(y)|$ stands for the absolute value of the first derivative of $v(y)$. Similarly,

$$f(x) = g(t(x))|t'(x)| = g(t(x)) = g(x+l) \quad \text{for} \quad a \leq x \leq b$$

From here, through the change of variable $y = x + l$:

$$\begin{aligned}\mu'_r(Y) &= \int_{a+l}^{b+l} y^r g(y)dy \\ &= \int_a^b (x+l)^r g(x+l)dx \\ &= \int_a^b (x+l)^r f(x)dx\end{aligned}$$

Particularly,

$$\mu'_1(Y) = l + \int_a^b x f(x)dx = \mu'_1(X) + l$$

and

$$\mu'_1(X) = \mu'_1(Y) - l$$

Finally,

$$\begin{aligned}\mu_r(Y) &= \xi(Y - \mu'_1(Y))^r \\ &= \int_{a+l}^{b+l} (y - \mu'_1(Y))^r g(y)dy \\ &= \int_a^b (x - \mu'_1(X))^r f(x)dx \\ &= \mu_r(X)\end{aligned}$$

It can be concluded that, under translations, the distribution moments remain unchanged. In particular, the variances and kurtosis remain equal.

Definition 2. let $\mathfrak{F}$ be the class of all symmetric distributions, bounded and definable on $\mathbb{R}$. It will be said that f and g are similar if g is a translation of f.

Theorem 2. *The similarity defined between elements of $\mathfrak{F}$ is an equivalence relation.*

This theorem, with inmediate proof, allows us to consider the symmetric, bounded, defined on $[-a, a]$ and centered at the origin distributions as representatives of the equivalence classes. Therefore, what can be said for one such distribution –related to its shape– is applicable to all the class members. In particular, these distributions satisfy Finucan's Principle. This fact, besides the partial order introduced by Van Zweet, allows us to extend the dominance range for symmetric distributions centered at the origin, as follows.

$$\cup\text{- Shaped Distrib} \quad \leq \quad \text{Uniform Distrib} \quad \leq \quad \cap\text{- Shaped Distrib}$$

And keeping in mind what was said for the kurtosis, it is also true that:

$$\gamma(\cup) \quad \leq \quad \gamma(\text{Uniform}) \quad \leq \quad \gamma(\cap)$$

4. Triangular Distributions

The symmetric triangular distributions with mean 0 play an important role in this work since they can be situated between the uniform distributions and the $\cup$ or $\cap$ – shaped distributions. These distributions may be defined, based on a parameter t which varies on the interval $[0, 1]$, by the following expressions:

1. Density function:

$$f(x) = \begin{cases} \frac{1}{a} - \frac{1-2t}{a^2}x, & \text{if} \quad -a \leq x \leq 0 \\ \frac{1}{a} + \frac{1-2t}{a^2}x, & \text{if} \quad 0 \leq x \leq a \\ 0, & \text{in other case} \end{cases}$$

2. Moments:

$$\mu_r = \mu'_r = \begin{cases} 2a^r\left[\frac{t}{r+1} - \frac{2t-1}{r+2}\right], & \text{if } r \text{ is even} \\ 0, & \text{if } r \text{ is odd} \end{cases}$$

3. Variance:

$$\sigma^2 = a^2\frac{3-2t}{6}$$

4. Kurtosis:

$$\gamma = \frac{12}{5}\frac{(5-4t)}{(3-2t)^2}$$

5. Cumulative distribution function:

$$F(x) = \begin{cases} 0, & \text{if } x \leq -a \\ \frac{t}{a}(x+a) - \frac{1-2t}{2a^2}(x^2-a^2), & \text{if } -a \leq x \leq 0 \\ \frac{1}{2} + \frac{t}{a}x + \frac{1-2t}{2a^2}x^2, & \text{if } 0 \leq x \leq a \\ 1, & \text{if } x \geq a \end{cases}$$

The above distributions can be classified in two groups, as follows:

a) Concave, when $t < \frac{1}{2}$

b) Convex, when $t > \frac{1}{2}$

The case $t = \frac{1}{2}$ corresponds to the uniform distribution.

The combination of Finucan's Principle and Van Zweet's ordering allows us to establish the following relation:

$$\gamma(\cup) \leq \gamma(\Delta\text{concave}) \leq \gamma(\text{Uniform}) \leq \gamma(\Delta\text{convex}) \leq \gamma(\cap)$$

From the simulations performed (see table) it can be concluded that there must exist an inverse relation (in the sense that while one increases the other decreases) between the kurtosis and the minimum sample size in order for it to be possible to use the normal approximation to estimate the distribution mean. Therefore, the

distributions that need a larger size are U−shaped, followed by concave triangular and then by uniform. According to this, the convex triangular distributions can be used to determine the minimum sample sizes to estimate the mean. These sizes usually are larger than the theoretic minimum established by the classical statistical literature which guarantees the estimation within a proposed confidence level.

5. Simulation And Results

The simulation, using Monte Carlo methods (Cfr [3] and [4]), was carried out developing an algorithm which can be summarized in the following steps: STEP 1. Fix n as a small value (30 is usually appropriate) STEP 2. Generate 1000 random samples of size n, from a symmetric triangular distribution determined by a value t in the interval $[0, 1]$ STEP 3. Construct, for each sample, the 95% confidence interval for the theoretic mean which is 0. This interval is given by:

$$[\bar{X} - z\frac{s}{\sqrt{n}}, \bar{X} + z\frac{s}{\sqrt{n}}]$$

where $\bar{X}$ is the sample mean, s^2 is the sample variance and z the quantile corresponding to the prefixed confidence level (1.96 in this case). STEP 4. Verify if the confidence interval contains the population mean and register a success in affirmative case. STEP 5. If the number of successes is greater than or equal to 950 (95% of the cases) n is considered appropriate. If it is not, n is increased by one unit and the process restarts at step 2.

TABLE I

SIMULATION RESULTS

t	a=0.2	a=1.0	a=2.0	a=3.0	Kurtosis
0.0	49	51	51	50	1.3333
0.1	48	49	50	50	1.4081
0.2	48	50	50	48	1.4911
0.3	47	50	50	51	1.5833
0.4	44	46	46	46	1.6859
0.5	41	42	43	44	1.8000
0.6	43	41	44	44	1.9259
0.7	42	40	42	43	2.0625
0.8	42	40	42	43	2.2040
0.9	41	41	42	43	2.3333
1.0	43	40	41	41	2.4000

The program execution yielded a series of values which may be interpreted as follows:

1. For bounded and convex symmetric distributions (bell - shaped), and even for the uniform, a sample of 44 items, randomly chosen, are enough to make use of the normal approximation for the mean estimation with a 95% confidence level.

2. The above value may be compared with $n = 30$, suggested value for normal or approximated normal distributions. In this case, however, the fact that the distribution is bounded has not been considered, as usually occurs for most situations in actual life.

BIBLIOGRAPHY

1. Balanda K. and MacGillivray H. L. (1988), *Kurtosis: A Critical Review*, The American Statistician Vol 42 No 2..
2. Hogg R. and Tanis E. (1989), *Probability and Statistical Inference*, Maxwell MacMillan, Singapur..
3. Ross S. (1991), *A Course in Simulation*, MacMillan Publishing Co., New York..
4. Rubinstein R. (1981), *Simulation and the Monte Carlo Method*, John Wiley, New York..