Journal of the Nigerian Statistical Association Vol. **28**, 2016, 74– 78

Generalizing Sample Size of Normally Distributed Samples using Generalized Exponential Power Distribution

T. Soyinka*
Federal Neuropsychiatric Hospital, Aro, Abeokuta, Nigeria

There are various sample size estimation formulas that are published in literature but has no adequate mathematical and statistical background. Many of such formula often assumed normal distribution that becomes unreliable most especially when observations are few. This study thus established sample size estimation formula from generalized exponential power distribution (GEPD) which has normal, Laplace and uniform distribution has its members. We employed an approximation to the incomplete gamma cumulative distribution function of the GEPD via series expansion to obtain the pivotal quantity from which the sample size of GEPD was derived. Application to sample size calculation from Likert scaled questionnaire was demonstrated.

Keywords: Likert scale; generalized exponential power distribution; pivotal quantity; sample size.

1. Introduction

According to Lindsey (1999) the random variable X is said to have univariate generalized exponential power distribution (GEPD) if

$$f(x;\mu,\sigma,\beta) = \frac{1}{\sigma\Gamma\left(1+\frac{1}{2\beta}\right)2^{1+\frac{1}{2\beta}}}\exp\left[-\frac{1}{2}\left|\frac{x-\mu}{\sigma}\right|^{2\beta}\right], \tag{1}$$

$-\infty < x < \infty; -\infty < \mu < \infty, \sigma > 0, \beta > 0$, where β is the shape parameter, μ and σ are location and scale parameters respectively.

If $\beta = 1/2$, (1) becomes a Laplace function. If $\beta = 1$, then (1) becomes a normal density and (1) approaches a uniform density as values of β increases beyond one towards infinity. However for $\beta < 1$ the distribution has heavier tails that is useful in providing robustness towards outliers (Gomez et al., 1998; Saralees, 2005).

Likewise the corresponding cumulative distribution function (CDF) for the GEPD is

$$F(x) = \frac{1}{2} + \frac{1}{2}sgn(x-\mu)\frac{\gamma\left[\frac{1}{2\beta}, \frac{1}{2}\left|\frac{x-\mu}{\sigma}\right|^{2\beta}\right]}{\Gamma\left(\frac{1}{2\beta}\right)}, \tag{2}$$

where γ is the upper incomplete gamma function. Simplifying (2) further using incomplete gamma expansion proposed by Takenaga (1966), Paris (2010) we obtain

$$F(x) = \frac{\left[\frac{1}{2}\right]^{\frac{1}{2\beta}+1}}{\Gamma\left(\frac{1}{2\beta}+1\right)}\left[\frac{\left|\frac{x-\mu}{\sigma}\right|}{0!} + \frac{\left(-\frac{1}{2}\right)\left|\frac{x-\mu}{\sigma}\right|^{2\beta+1}}{1!(2\beta+1)} + \frac{\left(\frac{1}{2}\right)^2\left|\frac{x-\mu}{\sigma}\right|^{4\beta+1}}{2!(4\beta+1)} + \frac{\left(-\frac{1}{2}\right)^3\left|\frac{x-\mu}{\sigma}\right|^{6\beta+1}}{3!(6\beta+1)} + \cdots\right] \tag{3}$$

*Corresponding author. Email: soyinkataiwo@yahoo.ie

which can be re-expressed as

$$F(x) = \frac{\left[\frac{1}{2}\right]^{\frac{1}{2\beta}+1}}{\Gamma\left(\frac{1}{2\beta}+1\right)} \sum_{r=0}^{\infty} \left[\frac{(-1)^r \left(\frac{1}{2}\right)^r \left|\frac{x-\mu}{\sigma}\right|^{2\beta r+1}}{r!(2\beta r+1)} \right] \tag{4}$$

in series form (Winitzki, 2003).

2. Pivotal Quantity for GEPD

2.1 *Definition*

Let $X_1, X_2, \cdots, X_n$ be a random variable each i.i.d from $f(x|\theta)$ and define $F(a|\theta) = \int_{-\infty}^{a} f(x|\theta)dx$; then a random variable $U = -2\ln F(X|\theta)$ has density $\left(\frac{1}{2}e^{-\frac{U}{2}} I_{(0,\infty)}(U)\right)$ which is a χ^2_2 density defined for every $F(X|\theta)$ over the uniform distribution interval $(0,1)$. Likewise the random variable $V = -2\ln[1 - F(X|\theta)]$ is χ^2_2 density. So if for every $i = 1, 2, \cdots, n$ we define $U_i = -2\ln F(X_i|\theta)$ then $U_1, U_2, \cdots, U_n$ are i.i.d. pivotal points each having χ^2_2 density.

Hence a sum across the pivotal points $PQ(X_i, \theta) = \sum_{i=1}^{n} U_i = -2\sum_{i=1}^{n} \ln F(X_i)$ has a χ^2_{2n} density and so is a pivotal quantity (PQ) for θ (Mood and Alexander, 1974; Zacks, 1981; Suhasini, 2010). Also $PQ_2(X_i, \theta) = \sum_{i=1}^{n} V_i = -2\sum_{i=1}^{n} \ln[1 - F(X_i)]$ is a pivotal quantity (PQ) for θ with χ^2_{2n} density.

The pivotal quantity (PQ) for GEPD is thus

$$PQ = -2\sum_{i=1}^{n} \ln \left(\frac{\left[\frac{1}{2}\right]^{\frac{1}{2\beta}+1}}{\Gamma\left(\frac{1}{2\beta}+1\right)} \sum_{r=0}^{\infty} \left[\frac{(-1)^r \left(\frac{1}{2}\right)^r \left|\frac{x-\mu}{\sigma}\right|^{2\beta r+1}}{r!(2\beta r+1)} \right] \right) \tag{5}$$

Note that equation (5) has odd powers of $\left|\frac{x_i-\mu}{\sigma}\right|$ when $\beta \geq 1\ \forall \beta \in \mathbb{Z}^+$ and all natural number powers when $\beta = \frac{1}{2}$. Hence equation (5) approximate to

$$PQ = \begin{cases} -2n\ln\frac{1}{\sqrt{2\pi}} - 2n\ln\left(\sum_{r=0}^{\infty}\frac{1}{2^r r!}\right) - 2\sum_{i=1}^{n}\ln\left(\sum_{r=0}^{\infty}\frac{(-1)^r\left|\frac{x_i-\mu}{\sigma}\right|^{2r+1}}{2r+1}\right) & \text{if } \beta \geq 1 \\ -2n\ln\frac{1}{4} - 2n\ln\left(\sum_{r=0}^{\infty}\frac{1}{2^r r!}\right) - 2\sum_{i=1}^{n}\ln\left(\sum_{r=0}^{\infty}\frac{\left|\frac{x_i-\mu}{\sigma}\right|^{r+1}}{r+1}\right) & \text{if } \beta = \frac{1}{2} \end{cases}$$

Concentrating on the part with powers of $\left|\frac{x_i-\mu}{\sigma}\right|$, $i = 1, 2, \cdots, n$, we obtain pivotal quantity for μ

$$Z_{i1} = \ln\left[1 + \left|\frac{x_i-\mu}{\sigma}\right|\right];\ -1 < \left|\frac{x_i-\mu}{\sigma}\right| \leq 1,\ i = 1, 2, \cdots, n \tag{6}$$

$$Z_{i2} = \tan\left|\frac{x_i-\mu}{\sigma}\right|;\ -\frac{\pi}{2} < \left|\frac{x_i-\mu}{\sigma}\right| < \frac{\pi}{2},\ i = 1, 2, \cdots, n. \tag{7}$$

Since the limits of $\left|\frac{x_i-\mu}{\sigma}\right|$ are independent of μ and σ then Z_{i1} and Z_{i2} are indeed pivotal quantity.

2.2 *Confidence interval for the location parameter (μ) with known scale parameter*

Considering equations (6) and (7), we obtain the confidence interval for the location parameter μ

$$P\left(q_1 < \ln\left[1 + \left|\frac{x_i - \mu}{\sigma}\right|\right] < q_2\right) = \gamma \tag{8}$$

where q_1 and q_2 are standard normal deviates at specified probability level $0 < \gamma < 1$ (Johnson and Wichern, 2006).

Suppose the observed values x_i's are sampling distribution of sample means (Jung et al., 2007). Then the confidence interval for the location parameter μ from the mean of its sample means is

$$P\left[\overline{X} - \frac{s}{\sqrt{n}}(e^{q_1} - 1) < \mu < \overline{X} + \frac{s}{\sqrt{n}}(e^{q_2} - 1)\right] = \gamma \tag{9}$$

where $s = \mathbb{E}(\sigma)$. Likewise from equation (7), the confidence interval for μ is

$$P\left[\overline{X} - \frac{s}{\sqrt{n}}\tan^{-1}|q_1| < \mu < \overline{X} + \frac{s}{\sqrt{n}}\tan^{-1}|q_2|\right] = \gamma \tag{10}$$

So constructing the 95% confidence interval for μ in the two cases for a unit scale we have

$$\overline{X} + \frac{0.859}{\sqrt{n}} < \mu < \overline{X} + \frac{6.099}{\sqrt{n}} \text{ and } \overline{X} - \frac{1.099}{\sqrt{n}} < \mu < \overline{X} + \frac{1.099}{\sqrt{n}}$$

respectively.

3. Sample Size Estimation Formula Required to Sample from GEPD

Making n the subject of the formula in equation (9) we obtain

$$n = \frac{(e^q - 1)^2 s^2}{\left|\overline{X} - \mu\right|^2} = \frac{(e^q - 1)^2 s^2}{E^2} \tag{11}$$

where E is the mean deviation. Assuming $q = Z_{\alpha/2} + Z_\beta$ at the level of type I (α) and type II (β) error (Gadbury et al. (2004)) then

$$n = \frac{(e^{Z_{\alpha/2}+Z_\beta} - 1)^2 s^2}{\left|\overline{X} - \mu\right|^2} = \frac{(e^{Z_{\alpha/2}+Z_\beta} - 1)^2 s^2}{E^2} \tag{12}$$

Note: $s = p(1 - p)$ can be substituted to obtain n if the only information available is the prevalence level from previous or similar study.

4. Samples Size Estimation Formula for Likert Scale Questionnaire

Depending on the scale of questionnaire under study, the above sample size formula can be adjusted to suit all cases of Likert scale measurements. If a Likert scale has options Yes/Neutral/No then its mean deviation is scaled over a unit standard scale. Also for

Likert scale Strongly Agree (SA), Agree (A), Neutral (N), Disagree (D), Strongly Disagree (SD) the mean deviation is a multiple of two unit standard scales. So for Likert scale measurements, as the options increases on either side of the divides, the multiplying factor (k) of the unit scale increases. So we have

$$n = \frac{(e^{Z_{\alpha/2}+Z_\beta} - 1)^2 s^2}{\left|\overline{X} - \mu\right|^2} = \frac{(e^{Z_{\alpha/2}+Z_\beta} - 1)^2 s^2}{(ks)^2} = \frac{(e^{Z_{\alpha/2}+Z_\beta} - 1)^2}{k^2}.$$

To accommodate for various errors in sample survey, we may double the sample size $N = 2n$.

Table 1: Sample size of Likert scaled questionnaire

$1-\alpha$	$Z_{\frac{\alpha}{2}}$	β	$Z_{\beta=0.2} = 0.85$		$Z_{\beta=0.1} = 1.282$	
			k=1	k=2	k=1	k=2
90%	1.645	$\beta \geq 1$	247.386	61.846	624.5661	156.1415
		$\beta = 1/2$	9291.301	2322.825	10121.069	2530.267
95%	1.96	$\beta \geq 1$	487.34	121.8348	1208.829	302.207
		$\beta = 1/2$	9915.357	2478.839	10616.442	2654.111
99%	2.576	$\beta \geq 1$	1770.527	442.63	4300.449	1075.112
		$\beta = 1/2$	10871.706	2717.926	11391.030	2847.758

5. Conclusion

This research work provided a better alternative to sample size calculation in observational study via a generalized exponential power distribution with flexible shape parameter as against the assumption of fixed shape parameter normal distribution. The results are easy to apply to any questionnaire with appropriate and statistically bound measurement of scales.

References

Gadbury, G.L., Page, G.P., Edwards, J., Kayo, T,. Prolla, T.A., Weindruch, R., Permana, P.A., Mountz, J.D., and Allison, D.B. (2004). Power and sample size estimation in high dimensional biology. *Statistical Methods in Medical Research*, 13:325–338.

Gomez, E., Gomez-Villegas, M.A., and Martin, J.M. (1998). A multivariate generalization of the exponential power family of distributions. *Communications in Statistics*

A27: 589–600.

Johnson, R.A. and Wichern, D.W. (2006). *Applied Multivariate Statistical Analysis.* Englewood Cliffs, NJ: Prentice-Hall, Inc.

Jung, S.H., Chow, S.C., and Chi, E.M. (2007). On sample size calculation based on propensity analysis in non-randomized trials. *Journal of Biopharmaceutical Statistics*, 17: 35–41.

Lindsey J.K. (1999). Multivariate elliptically contoured distributions for repeated measurements. *Biometrics* 55: 1277–1280.

Mood, A. M., Graybill, F. A. and Boes, D. C. (1974). Introduction to the Theory of Statistics (McGraw-Hi1l Series in Probability and Statistics). ISBN 0-07-042864-6.

Nadarajah, S. (2005). A generalized normal distribution. *Journal of Applied Statistics* 32(7): 685–694. DOI:10.1080/02664760500079464.

Paris, R.B. (2010). *Incomplete Gamma Function.* NIST Handbook of Mathematical Functions, Cambridge University press.

Stacy, E.W. (1962). A generalization of the gamma distribution. *Annals of Mathematical Statistics*, 33(3): 1187–1192.

Suhasini, S.R. (2010). Advanced Statistical Inference. suhasini.subbarao@stat.tamu.edu.

Takenaga, R. (1966). On the Evaluation of the Incomplete Gamma Function. *Math. Comp.*, 20 (96): 606–610.

Winitzki, S. (2003). Computing the incomplete gamma function to arbitrary precision. *Lecture Notes Comp. Sci.* 2667: 790–798.

Zacks, S. (1981). *Parametric Statistical Inference Basic Theory and Modern Approaches.* Pergamon Press State University of New York Binghamton.